

Proceedings of the 2009 IMS Conference for New Researchers in Statistics and Probability

Preface

The 2009 Institute for Mathematical Statistics (IMS) New Researchers Conference (NRC) was held at John Hopkins University from July 28th to July 31st, 2009, prior to the Joint Statistical Meetings in Washington D.C. The New Researchers Conference gives statisticians and probabilists the opportunity to meet their colleagues in a more intimate setting than the usual annual conferences provide and discuss their research. The NRC can serve as a weathervane pointing out the direction that the field is heading in. In 2009, many of the participants discussed spatiotemporal modeling and high dimensional inference. The meeting also indicated the ever increasing presence of Bayesian inference and modeling. This year, a first for NRC, a collection of papers associated with the conference was introduced in order to serve as proceedings of the IMS-NRC conference. This collection of scientific papers aims to document the presentations given at the meeting. Participants were given the opportunity to submit short papers (or notes) on the material they presented with the hope that they will be eventually submitted to a journal in statistics, probability or allied fields.

The IMS NRC organizing committee would like to take this opportunity to thank the sponsors of the event (IMS, NSA, NSF, NIH). We thank our invited speakers Nanny Wermuth from University of Gothenburg, Bernard Silverman from Oxford University and Tom Louis from Johns Hopkins University. We are grateful to the Department of Biostatistics in the Bloomberg School of Public Health at Johns Hopkins University for hosting the meeting, and particularly its chair Karen Bandeen-Roche for her participation and advice. We would like to thank the Biostatistics Branch in the Division of Cancer Epidemiology and Genetics at the National Cancer Institute for also hosting the meeting, and Nilanjan Chatterjee and Hormuzd Katki for their assistance organizing the meeting on their campus. We also thank Kate Schmidt, Linda Zenner and Megan Schlick at the University of Minnesota for their assistance administering the meeting.

The organizing committee would also like to thank the participants for making this NRC one of the most successful to date.

Editors

Tracy Bergemann (University of Minnesota), Bala Rajaratnam (Stanford University)

Schedule of Presentations

IMS New Researchers Conference Program July 28th - July 31st, 2009

Arrival July 27th

2:00--4:00pm Check into Charles Commons
3301 N. Charles St., Baltimore, MD 21218
<http://www.jhu.edu/hds/oncampus/buildings.html>

Scientific Sessions will take place in Maryland Hall 110
<http://www.jhu.edu/tour/maryland.html>
This can be located on the campus map <http://www.jhu.edu/tour/map.html>

Schedule July 28th -- Maryland Hall 110

- 8:00-8:45 Breakfast at Nolan's Cafe
- 8:45-9:00 Introductory Remarks
- **9:00-10:20 Scientific Session on Spatial-temporal data**
 - Farouk Nathoo, University of Victoria, Joint Spatial Modelling of Recurrent Infection and Growth with Processes Under Intermittent Observation
 - Jing Zhang, Miami University, Zero-inflated Bayesian spatial models with repeated measurements
 - Pál Rakonczai, Eötvös Loránd University, Hungary, Modeling Multivariate Extremes for Wind Speed Data
 - Debashis Mondal, University of Chicago, De Wijs Process and Modeling Disease Risk
 - Guilherme Rocha, Indiana University, Monitoring Civil Structures Using Restricted Autoregressive Models and Wireless Sensor Networks (WSNs)
- 10:20-10:40 Break
- **10:40-12:00 Scientific Session on Spatial-temporal data**
 - Lily Wang, University of Georgia, Spline-Backfitted Kernel Smoothing of Additive Models in Time Series
 - Jin-Hong Park, College of Charleston, Central Mean Subspace in Time Series
 - Peng Zhang, University of Alberta, Joint Mean-covariance Modeling with Nonlinear Random Mean Models for Longitudinal Data Analysis

- Rongning Wu, Baruch College, A Negative Binomial Model for Time Series of Counts
- Olu Awosoga, Western Michigan University, Meta Analyses of Multiple Baseline Time Series Design Intervention Models for Dependent and Independent Series
- 12:00-1:45 Lunch at Nolan's Cafe
- **1:45-2:50 Scientific Session on Bayesian Inference**
 - Kshitij Deepak Khare and Bala Rajaratnam, Stanford University, Inference in Gaussian Covariance Graph Models
 - Weixin Yao, Kansas State University, Bayesian Mixture Labeling by Highest Posterior Density
 - James Flegal, University of California, Riverside, Markov Chain Monte Carlo: Can We Trust the Third Significant Figure?
- **2:50-3:15 Break**
- **3:15-4:30 Tweedie Award Speaker**

Jiashun Jin, Department of Statistics, Carnegie Mellon University

Higher Criticism Thresholding: Optimal Feature Selection when Useful Features are Rare and Weak

- 5:00-- ? Social outing in Baltimore -- take the Hopkins shuttle bus to Peabody
http://www.parking.jhu.edu/shuttles_jhmi_homewood.html

Restaurants in Mt Vernon:

- Helmand at 806 N Charles St, \verb+<http://www.helmand.com>+, Afghani food
- Indigma at 802 N Charles St, \verb+<http://www.indigmarestaurant.com>+, Indian food
- My Thai at 800 N Charles St, \verb+<http://www.mythaibaltimore.com>+, Thai food

Schedule July 29th -- Maryland Hall 110

- 8:00-8:45 Breakfast at Nolan's Cafe
- 8:45-9:00 Introductory Remarks
- **9:00-10:20 Scientific Session on Experimental Design**
 - Tirthankar Dasgupta, Harvard University, D-Optimal and Sequential Designs for Estimating Parameters of the Linear-Exponential Growth Curve of Nanowires
 - Ying Hung, Rutgers, the State University of New Jersey, Design and Analysis of Computer Experiments with Branching and Nested Factors
 - Edgard Maboudou-Tchao, University of Central Florida, Some Theoretical properties of the Multivariate Exponentially Weighted Moving Covariance Matrix
 - Vincent Agboto, Meharry Medical College, A Bayesian Approach to Model Robust Designs
 - Jingchen Liu, Columbia University, Statistics Can Lie But Can Also Correct for Lies: Reducing Response Bias in NLAAS via Bayesian Imputation
- 10:20-10:40 Break

- **10:40-11:45 Scientific Session on Count Data**
 - ChengYong Tang, National University of Singapore, Local Post-Stratification in Dual System Accuracy and Coverage Evaluation for US Census
 - Wanhua Su, University of Alberta, Statistical Inference on Recall, Precision and Average Precision
 - Hyokyung Hong, Baruch College, Prediction of Conditional Quantiles Based on a New Ordinal Regression Model
 - Xiaofeng Wang, Case Western Reserve University, Spatial models for replicated regional count data in neuroscience
- 12:00-1:30 Lunch at Nolan's Cafe
- **1:30-2:35 Scientific Session on Model Selection**
 - Sonja Greven, Johns Hopkins University, On the Behavior of Marginal and Conditional Akaike Information Criteria in Linear Mixed Models
 - Qin Wang, University of Georgia, Sufficient Dimension Reduction and Variable Selection: A Localized Approach
 - Julia Salzman, Stanford University, Detecting Multivariate Dependencies in High Dimensional Data
 - Hugh Miller, University of Melbourne, Local polynomial regression incorporating variable selection
- 2:35-3:00 Break
- **3:00--4:00 Mentoring Panel**
- **5:00--7:30 Dinner Banquet with Speaker Karen Bandeen-Roche, "Latent Variables and Your Future", Salon C in Charles Commons**

NIH/NCI schedule July 30th -- NIH campus

- 6:45-7:00 Board the Broadway Shuttle bus in front of Charles Commons
- 7:00-8:00 Commute to Bethesda
- 8:00-8:30 Arrive at NIH, pass through security (everyone will require a driver's license or passport)
- 8:30-8:45 Coffee and Breakfast
- 8:45-9:00 Introduction by Nilanjan Chatterjee, Chief of the DCEG Biostatistics Branch
- **9:00-9:50 Scientific Session on Biostatistics**
 - Chiung-Yu Huang, Mathematical Statistician, Biostatistics Research Branch, NIAID, NIH, Analysis for recurrent event data with informative censoring
 - Ju Hyun Park, Visiting Fellow, Biostatistics Branch, DCEG, NCI, Evaluation of power and Risk prediction utility of future genome-wide association studies

- 9:50-10:10 Break
- **10:10-11:15**
 - Hormuzd Katki, Tenure-Track P.I., Biostatistics Branch, DCEG, NCI, Insights into p-values and Bayes Factors from False Positive and False Negative Bayes Factors
 - Norou Diawara, Old Dominion University, Mixture of Bivariate Exponential Distributions
 - Rajeshwari Sundaram, Tenure Track PI, Biostatistics and Bioinformatics Branch, DESPR, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Statistical challenges in modeling human fecundity
- **11:15-12:00 NIH Branch Chief Panel:**
 1. Paul Albert, Biostatistics and Bioinformatics Branch, NICHD
 2. Eric (Rocky) Feuer, Statistical Research and Applications Branch, NCI
 3. Gang Zheng, Office of Biostatistics Research, NHLBI
 4. Edward Korn, Biometric Research Branch, NCI
- **12:00-1:30 Lunch with Speaker Kai Yu, Tenure-Track Principal Investigator, Biostatistics Branch, DCEG, NCI, "Statistical Challenges in Gene Mapping"**
- **1:40-2:45 Scientific Session on Statistical Genetics**
 - Huilin Li, Research fellow, DCEG, NCI, Finding SNP Associations with a Secondary Phenotype in Genetic Association Studies
 - Dawei Liu, University of Iowa, Estimation and Testing for the Effect of a Genetic Pathway on a Disease Outcome Using Logistic Kernel Machine Regression via Logistic Mixed Models
 - Jason Wilson, Biola University, Evidence that Oligonucleotide Expression Values Are Not Normally Distributed
 - Arabin Kumar Dey, IIT Kanpur, Discriminating between two Bivariate Survival Models: Bivariate Weibull and Bivariate Generalized Exponential distribution
- 2:45-3:10 Break
- **3:10-4:00**
 - Samsiddhi Bhattacharjee, Research Fellow, Biostatistics Branch, DCEG, NCI, Robust Detection of Gene-Gene Interactions in Presence of Hidden Stratification.
 - Nak-Kyeong Kim, Old Dominion University, Finding Sequence Motifs with Bayesian Models Incorporating Positional Information
 - Qunhua Li, University of California, Berkeley, Measuring the consistency of high-throughput biological experiments
- **4:00-4:45 Panel of NIH Grant Agencies**
 1. Michelle Dunn, Statistician Research and Applications Branch, NCI: "Grant Opportunities for New Researchers"
 2. Misrak Gezmu, Biostatistics Research Branch, NIAID: "NIH grant mechanisms and funding process"
- 5:00-? Head to downtown Bethesda for dinner
<http://www.downtownbethesda.com/guide/dining.php>

Schedule July 31st -- Maryland Hall 110

- 8:00-8:45 Breakfast at Nolan's Cafe
- 8:45-9:00 Introductory Remarks
- **9:00-10:05 Scientific Session 9**
 - Jiguo Cao, Simon Fraser University, Statistical Inference for Dynamic models with the Generalized Profiling Method
 - Zi Jin, University of Toronto, Efficiency of Composite Likelihoods
 - Bodhisattva Sen, Columbia University, Bootstrap in some Non-standard Problems
 - Yuefeng Wu, North Carolina State University, Posterior Consistency for some Semi-parametric Problems
- 10:05-10:25 Break
- **10:25-11:45 Scientific Session 10**
 - Ahmad Yasamin, Statistical and Applied Mathematical Sciences Institute, On Existence of the Maximum Likelihood Estimator for Gaussian Graphical Models
 - Yanping Xia, Southeast Missouri State University, Generalized Variable Approach for Correlation Analysis
 - Jason Morton, Stanford University, Algebraic Models for Multilinear Dependence
 - David King, Arizona State University, Topics in Functional Canonical Correlation
 - Andrada Ivanescu, East Carolina University, Adaptive Inference for Sparse Signals in Functional Data
- **12:00-2:00 Lunch Banquet with Speaker Nanny Wermuth, "The development of graphical Markov models: some history and some personal experiences."** Salon B in Charles Commons
- **2:00-3:30 Journal Panel -- Maryland Hall 110**
Bernard Silverman - Annals of Statistics ed., Thomas Louis - Biometrics ed.

Papers (by alphabetical order)

1. Meta Analyses of Multiple Baseline Time Series Design Intervention Models for Dependent and Independent Series
Oluwagbohunmi Awosoga, Joseph McKean, & Bradley Huitema
2. Notes on Bivariate Exponential Mixture for Heterogeneous Survival Data
Norou Diawara
3. On the Behavior of Marginal and Conditional Akaike Information Criteria in Linear Mixed Models
Sonja Greven & Thomas Kneib
4. Priors for sparse covariance models
Kshitij Khare & and Bala Rajaratnam
5. Exploratory Analysis of Multivariate Dependencies in High Dimensional Data
Kshitij Khare, Bala Rajaratnam & Julia Salzman
6. A Note on Congruence Relations for Second-Order Processes
David King
7. Multivariate Control Chart for Monitoring the Covariance Matrix
Edgard M. Maboudou-Tchao & Douglas M. Hawkins
8. A Note on the Probability of Correct Selection for Large k Populations, with Application to Microarray Data
Jason Wilson
9. Labelling Switching for Bayesian Mixture Models
Weixin Yao & Bruce Lindsay
10. On the Existence of the maximum likelihood estimator for Gaussian graphical models and a new graph invariant
Ahmad S. Yasamin
11. Nonlinear Joint Mean-covariance Modeling for Longitudinal Data Analysis
Peng Zhang & Peter X.-K. Song

META ANALYSES OF MULTIPLE BASELINE TIME SERIES DESIGN INTERVENTION MODELS FOR DEPENDENT AND INDEPENDENT SERIES

Oluwagbohunmi Awosoga, Joseph McKean, & Bradley Huitema

University of Lethbridge, Lethbridge, Alberta, Canada & Western Michigan University, Kalamazoo, Michigan, USA

Abstract : In this work we develop a traditional meta-type analysis for multiple baseline series. Robust methodology for multiple baseline series is also developed. The procedures are almost efficient as the traditional method on “good” data and are generally much more efficient on data containing outliers. The diagnostic procedures for the analysis of these data are also developed. The design matrices provided for the two-phase (AB) design allow for change in level and change in slope between each phase and the subsequent phase. Similarly, our methodology can be extended to more than two Phased time series design intervention models. Our parametric procedures are based on least squares (LS) estimation. The robust procedures are similar to the parametric procedures except another norm rather than the Euclidean norm is used. Illustrative examples are discussed. A Monte Carlo study of the methods are provided. The study investigates the validity of the procedures and power comparisons between the parametric and robust methods.

Key Words : Meta Analysis, Level Change, AB-Design, Autocorrelation, Double Bootstrap, Effect Size, Contaminated Normal distribution

1 Introduction

This work evaluates a proposed meta-analytic procedure for the analysis of multiple baseline series (Huitema, 2004), comparing it with a robust analog. Diagnostic procedures for the analysis of these designs are also developed. This presentation focuses on new approaches for the meta analyses of multiple baseline (series) AB-type designs.

Single subject designs are of great interest to researchers in the fields of behavioral sciences, environmental sciences, economics, education, and medicine. They often employ interrupted time-series designs to determine the effectiveness and efficiency of several interventions in both clinical and natural settings. Many researchers in the behavioral sciences concur that multiple baseline time series intervention designs provide a strong basis for causal conclusions (Koehler and Levin, 1998).

Often, researchers are interested in seeing whether the intervention produces an immediate and persistent change in level or slope. As recommended by Huitema and McKean (1998), we adopt the following model (H-M four parameter model):

$$Y_t = \beta_0 + \beta_1 T_t + \beta_2 D_t + \beta_3 [T_t - (n_1 + 1)] D_t + \epsilon_t \quad (1)$$

where at time t , Y_t is the dependent variable score, T_t is the value of the measurement occasion variable (usually t), D_t is the value of the level change dummy variable D (0 for first phase and 1 for second), β_0 is the regression intercept, β_1 through β_3 are the process partial regression coefficients, and ϵ_t is the process error.

This work studies the behavior of new parametric and robust approaches being introduced with specific focus on the following four scenarios: (1) independence within series and between two or more baselines; (2) dependence within series, but independence between two or more baselines; (3) independence within series, but dependence between two or more baselines; and (4) dependence within series, and between two or more baselines.

2 Methodology

In Awosoga (2009), we present details of our meta analyses for each of the four scenarios discussed above. Our parametric analysis is based on the ordinary least squares (OLS) fit while the robust meta analysis is based on the Wilcoxon rank-based fit described in Chapter 3 of Hettmansperger and McKean (1998). Upon fitting a series, residual diagnostics are used to check the quality of fit. In particular, the Durbin-Watson (DW) test and a simple, but powerful, test proposed by Huitema and McKean (2000b) (HM test) are used to check for autocorrelation in the errors. If autocorrelation is not detected, we compute the the meta analyses described in the next section. If autocorrelation is present, we recommend the double bootstrap procedure of McKnight et al. (2000) for a second fit, which we then use in the meta analyses; see Awosoga (2009) for details.

3 Algorithm for the New Meta Analytic Procedures

Suppose a data set is given in three columns: the first column consists of the baseline series number (*i.e.*, $1, 2, \dots, J$), second column represents the different phases in each series (*i.e.* phase 1 is pre-intervention and phase 2 is post-intervention), and the third column contains the response variables (*i.e.* the observed data). For each series, let y_1 = pre-intervention response variables, y_2 = post-intervention response variables, n_1 = length(y_1), n_2 = length(y_2), and $n = n_1 + n_2$ respectively.

Table 1: General Form of Summary Table for LC (/ SC / RLC /RSC) Preliminary Output

Series	LC	SE(LC)	TS(LC)	Pvallc	z (LC)	Weight	Effect Size
1	LC_1	SE_1	$t(LC_1)$	$P(LC_1)$	$z(LC_1)$	W_1	ES_1
2	LC_2	SE_2	$t(LC_2)$	$P(LC_2)$	$z(LC_2)$	W_2	ES_2
3
.
.
.
J	LC_J	SE_J	$t(LC_J)$	$P(LC_J)$	$z(LC_J)$	W_J	ES_J

Note: Same format goes for SC (except that we omit Effect Size on SC summary) and Robust (RLC and RSC) Summary Table.

3.1 The Meta Analytic Components

In addition to the preliminary output summary table discussed above, we provide the user with the meta analytic components needed for an overall decision making. This include the following computations (each gives a weighted average of all statistics):

3.1.1 Overall-test statistic for the LC

$$z_{overall} = \frac{\sum_{j=1}^J z_j}{\sqrt{J}}. \quad (2)$$

3.1.2 The weighted overall LC statistic

$$Overall \ Level \ Change = \frac{\sum_{j=1}^J [\frac{1}{\hat{\sigma}_{e_j}^2} (LCCoeff_j)]}{\sum_{j=1}^J \frac{1}{\hat{\sigma}_{e_j}^2}}. \quad (3)$$

3.1.3 The overall standardized effect size

$$Overall \ Standardized \ Effect = \frac{\sum_{j=1}^J \left[\frac{LCCoeff_j}{\sqrt{MSResid_j}} \right]}{J} \quad (4)$$

3.1.4 Test on the homogeneity of effects across units

$$Pair-wise \ Homogeneity \ Test = \frac{z_j - z_i}{\sqrt{2}}. \quad (5)$$

Extensive computer simulation studies were carried out (using R software and FORTRAN codes) over a variety of sample sizes and error distributions. Results were compared under each scenario stated above. These investigations are briefly summarized in the next section; see Awosoga (2009) for a detailed summary.

4 Conclusion & Future Study

Our new parametric method and its robust analog are recommended for scenario 1, respectively, when series are without and with outliers. The double bootstrap method discussed by McKnight et al. (2000) in conjunction with our meta analysis summary table should be used for the case of independence between series but dependence within series (scenario 2). For scenario 3, we recommend the correlation test (CT) or JR as discussed in Awosoga (2009). Finally, dependence between series and dependence within series is handled using a combination of paired difference and double bootstrap method. As an extension of this work, we will develop a R package, which can generate appropriate design matrices for individual users, combined with current program codes for multiple baseline time-series design intervention models. We plan to extend this work to more than two phases multiple baseline designs i.e. reversal designs such as ABA, ABAB, and ABABA designs.

BIBLIOGRAPHY

- Awosoga, O.A. (2009), "Meta analyses of multiple baseline time series design intervention models for dependent and independent series," Unpublished PhD Dissertation, Western Michigan University.
- Hettmansperger, T. P., & McKean, J. W. (1998), "Robust Nonparametric Statistical Methods," *Kendalls Library of Statistics 5, Great Britain, Arnold*.
- Huitema, B. E. (2004b), "Anaysis of interrupted time - series experiments using ITSE: A critique," *Understanding Statistics: Statistical Issues in Psychology, Education, & the Social Sciences*, 3, 27–46.
- Huitema, B. E., & McKean, J. W. (1998), "Irrelevant autocorrelation in least - squares intervention models," *Psychological Methods*, 3, 104–116.
- Huitema, B. E., & McKean, J. W. (2000a), "Design specification issues in time - series intervention models," *Education & Psychological Measurement*, 60, 38–58.
- Huitema, B. E., & McKean, J. W. (2000b), "A simple and powerful test for autocorrelated errors in OLS intervention models," *Psychological Reports*, 87, 3–20.
- Koehler, M. J., & Levin, J. R. (2000), "RegRand: Statistical software for the multiple - baseline design," *Behavior Research Methods, Instruments & Computers*, 32, 2, 367–371.
- Kloke, J. D., McKean, J. W., & Rasid, M. M. (2009), "Rank - Based Estimation and Associated Inferences for Linear Models with Clustered Correlated Errors," *Journal of the American Statistical Association*, 104, 485, 384–390.
- McKnight, S., McKean, J.W. & Huitema, B.E. (2000), "A double bootstrap method to analyze linear models with autoregressive error terms," *psychological Methods*, 5, 87–101.

NOTES ON BIVARIATE EXPONENTIAL MIXTURE FOR HETEROGENEOUS SURVIVAL DATA

Norou Diawara

Old Dominion University, Norfolk, Virginia, USA

Abstract : The exponential distribution is one of the most used type of distribution because of its importance in many lifetime applications and its properties. Its bivariate form based on a linear relationship is useful in introducing relationship. Simply used, there can be limitations especially for a heterogeneous type population. We present a mixture form and describe numerous desirable properties of the associated parameters and estimate the components of the mixture. We include the presence of covariate information, a special case of which was considered by Marshall and Olkin.

Key Words : Bivariate exponential, Mixture, Dirac delta, Maximum likelihood, EM algorithm

1 Introduction

Survival time of patients depends on the multiple diseases considered and is of a growing interest in the medical, engineering and statistical sciences to mention a few areas. The form we consider here captures the property of nonzero probability of simultaneous occurrence, from Marshall and Olkin (1967) [9]. Carpenter et al (2006) [2] presented a model for a parallel system with a bivariate distribution and derived some associated properties. More specifically, for f_1 and f_2 being the marginal densities of two random variables X_1 and X_2 in such a system, the joint distribution of (X_1, X_2) is given by:

$$g(x_1, x_2) = pf_1(x_1)\delta(x_2 - ax_1) + (1 - p)f_1(x_1)f_2(x_2 - ax_1)I(x_2 > ax_1), \quad (1)$$

where

- X_1, X_2 are two survival distributions related as $X_2 = aX_1 + Y$, with a a nonnegative fixed constant, and Y an unknown random variable.
- $p = P(X_2 = aX_1)$ is the probability of proportional occurrence.
- and $\delta(x)$ refers to the Dirac delta function, i.e. $\delta(t) = 0$, if $t \neq 0$, and $\int_{-\infty}^{\infty} \delta(t)dt = 1$. See Abramowitz and Stegun (1972) [1] for more details.

However, the use of one single distribution can have limitations especially for heterogeneous populations. When observations are from a heterogeneous population, implementation of the mixture adds a lot more choice and flexibility to the model. McLachlan and Peel (2000) [11] and Redner and Walker (1984) [12] give several motivations for the use of mixture densities. The exponential is very useful in many lifetime applications. Letting $X = (X_1, X_2)'$ be the bivariate form of the distribution, our goal is to study its exponential mixture form. We use finite classical mixture modelling via a modified version of the expectation-maximization (EM) algorithm from Dempster et al (1977) [3] and McLachlan and Peel (2000) [11]. Estimators of the parameters associated with the mixture of bivariate exponential distributions are introduced. These results and their applications will extend to situations in many other cases.

2 The Mixture of Bivariate Exponential

The exponential distribution is quite well known and has extensive use. For the random variable X with exponential distribution, the associated probability density function is defined with respect to its scale parameter λ , by:

$$f(x) = f_X(x) = \lambda e^{-\lambda x}, \quad \text{for } x > 0, \text{ and } \lambda > 0.$$

Based on (1), and on X_1 and X_2 each possessing an exponential with scale parameters λ_1 and λ_2 , respectively, the random vector $X = (X_1, X_2)'$ is said to have a bivariate exponential distribution, where $X_2 = aX_1 + Y$, for some fixed nonnegative constant a and an unknown random variable Y . Its density function $g(x|\theta)$ is given as in (1). Hence, based on a random sample of size N , $N \in \mathbf{N}$, the density function of $X_i = (X_{i1}, X_{i2})'$, $i = 1, \dots, N$, is given by:

$$\begin{aligned} g(x_i|\theta) &= g(x_{i1}, x_{i2}|\theta) = pf_{i1}(x_{i1})\delta(x_{i2} - ax_{i1}) \\ &\quad + (1-p)f_{i1}(x_{i1})f_{i2}(x_{i2} - ax_{i1})I(x_{i2} > ax_{i1}), \quad i = 1, \dots, N. \end{aligned}$$

Since we do not know from which part of the homogeneous population X_i originates, the marginal density is in the form of a mixture, and we say that g is a finite mixture of M component densities if it is given, as in Redner and Walker (1984) [12], by:

$$g(x_i|\theta) = \sum_{j=1}^M \alpha_j \mathbf{g}_j(\mathbf{x}_i|\theta), \quad \text{for } \mathbf{i} = \mathbf{1}, \dots, \mathbf{N}, \quad \text{with} \quad (2)$$

(i) $\theta = (\theta_1, \theta_2, \dots, \theta_M)$, with $\theta_j = (\lambda_{1j}, \lambda_{2j}, p_j)$ for $j = 1, \dots, M$.

(ii) The nonnegative coefficients α_j , $j = 1, \dots, M$, are such that $\sum_{j=1}^M \alpha_j = 1$.

From this model, we wish to estimate the parameters of the mixture density given in (2) under conditional expectation and maximization or E-M steps.

3 Results of the Parameters Inference

Consider a random sample of size N of bivariate vectors that are related as in (1), $\mathbf{x} = (x_1, \dots, x_N)$,

$$\hat{\alpha}_{ij} = \frac{\beta_j \lambda_{1j} e^{-\lambda_{1j} x_{i1}} \delta(x_{i2} - a_j x_{i1}) + \bar{\beta}_j \lambda_{1j} \lambda_{2j} e^{-(\lambda_{1j} - a_j \lambda_{2j}) x_{i1}} e^{-\lambda_{2j} x_{i2}} I(x_{i2} > a_j x_{i1})}{\sum_{k=1}^M \beta_k \lambda_{1j} e^{-\lambda_{1j} x_{i1}} \delta(x_{i2} - a_k x_{i1}) + \sum_{k=1}^M \bar{\beta}_k \lambda_{1j} \lambda_{2j} e^{-(\lambda_{1j} - a_k \lambda_{2j}) x_{i1}} e^{-\lambda_{2j} x_{i2}} I(x_{i2} > a_k x_{i1})},$$

with $\lambda_{lj} = \lambda_{lj}^{(t)}$, $l = 1, 2$, $1 \leq i \leq N$ and $1 \leq j \leq M$, and $\beta_j = \alpha_j p_j$ and $\bar{\beta}_j = \alpha_j (1 - p_j)$, $1 \leq j \leq M$.

Hence,

$$\hat{\alpha}_j^{(t+1)} = \frac{1}{N} \sum_{i=1}^N \hat{\alpha}_{ij}(\theta^{(t)}), \quad \mathbf{1} \leq \mathbf{j} \leq \mathbf{M}, \quad \mathbf{t} \geq \mathbf{0}.$$

We assume that we can possibly observe right censored lifetime data, and let

$$r_i = \begin{cases} 1 & \text{if } x_{i2} = a_j x_{i1} \text{ for some } j = 1, \dots, M, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

Hence,

$$\hat{\lambda}_{2j} = \frac{\sum_i \alpha_{ij}}{a_j \sum_i \alpha_{ij} x_{1i} + \sum_i \alpha_{ij} (1 - r_i) z_i}, \quad (4)$$

$$\text{and } \hat{\lambda}_{1j} = \frac{a_j \sum_i \alpha_{ij}}{a_j \sum_i \alpha_{ij} x_{1i} + \sum_i \alpha_{ij} (1 - r_i) z_i} + \frac{\sum_i \alpha_{ij} (1 - r_i)}{\sum_i \alpha_{ij} x_{1i}}. \quad (5)$$

Both second order partial derivatives of the likelihood evaluated at $(\lambda_{1j}, \lambda_{2j}) = (\hat{\lambda}_{1j}, \hat{\lambda}_{2j})$ are negative.

And the Hessian matrix is given by:

$$H(\lambda_{1j}, \lambda_{2j}) = \begin{pmatrix} -\frac{\sum_i \alpha_{ij} (1 - r_i)}{(\lambda_{1j} - a_j \lambda_{2j})^2} & a_j \frac{\sum_i \alpha_{ij} (1 - r_i)}{(\lambda_{1j} - a_j \lambda_{2j})^2} \\ a_j \frac{\sum_i \alpha_{ij} (1 - r_i)}{(\lambda_{1j} - a_j \lambda_{2j})^2} & -\frac{\sum_i \alpha_{ij}}{\lambda_{2j}^2} - a_j^2 \frac{\sum_i \alpha_{ij} (1 - r_i)}{(\lambda_{1j} - a_j \lambda_{2j})^2} \end{pmatrix}.$$

Hence,

$$\det(H(\lambda_{1j}, \lambda_{2j})) = \frac{\sum_i \alpha_{ij} (1 - r_i) \sum_i \alpha_{ij}}{\lambda_{2j}^2 (\lambda_{1j} - a_j \lambda_{2j})^2} > 0.$$

Therefore, $(\hat{\lambda}_{1j}, \hat{\lambda}_{2j})$ as in (5) and (4) is mle for the distribution. It can be seen that (4) and (5) are generalizations of the estimates obtained from Carpenter et al (2006) [2], where the mixing terms α_{ij} would be equal to 1. In fact, (5) and (4) would reduce respectively to:

$$\hat{\lambda}_2 = \frac{n}{a \sum_i x_{1i} + \sum_i z_i} = \frac{1}{\bar{x}_2} \quad \text{and} \quad \hat{\lambda}_1 = \frac{1}{\bar{x}_2} + \frac{n - k}{n \bar{x}_1}.$$

4 Conclusion

We have proposed a mixture of bivariate exponential distributions with a closed form expression of its density which complements those in Kotz et al (2000) [8], Johnson and Kotz (1970) [6] and Johnson and Wichern (1998) [7]. The Marshall-Olkin property of proportional occurrence is captured. We have included a flexible dependence as in Mathai and Moschopoulos (1992) [10], Iyer et al (2004) [4] and in Carpenter et al (2006) [2]. Estimation of parameters is completed using a classical mixture modelling technique via the EM algorithm. The computational cost is low compared to alternative methods. The results are quite interpretable, generalize outcomes previously obtained in the literature, and are very attractive and desirable according to criterion in Joe (1997) [5].

References

- [1] Abramowitz, M and Stegun, I.A. (1972), *Handbook of Mathematical Functions*, NY.
- [2] Carpenter, M., Diawara, N. and Han, Yi, (2006), *A New Class of Bivariate Weibull Survival Distributions*, J. Math. & Managt. Sciences, Vol. 26 (1 & 2), pp. 164-184.
- [3] Dempster, A., Laird, N. and Rubin, D. (1977), *Maximum Likelihood from Incomplete Data via the EM algorithm*, J. Royal Statistical Society, Series B, Vol 39 (1), pp 1-38.
- [4] Iyer, Srikanth K. and Manjunath, D. (2004), *Correlated Bivariate Sequence for queuing and reliability applications*, Communications in Statistics, Vol. 33, pp. 331-350.
- [5] Joe, H. (1997), *Multivariate Models and Dependence Concepts*, Chapman & Hall.
- [6] Johnson, N. and Kotz, S. (1970), *Continuous Univariate Distributions-1* Wiley, NY.
- [7] Johnson, R.A. and Wichern, D.W. (1998), *Applied Multivariate Statistical Analysis*, Prentice Hall, 4th edition.
- [8] Kotz, S., Balakrishnan, N. and Johnson, N. (2000), *Continuous Multivariate Distributions*, Volume 1 Wiley Series in Probability and Statistics.
- [9] Marshall, A.W. and Olkin, I. (1967), *A Multivariate Exponential Distribution*, J. Amer. Stat. Assoc., 63: pp. 30-44.
- [10] Mathai, A.M. and Moschopoulos, P.G. (1992), *A Form of Multivariate Gamma Distribution*, Annals of the Institute of Statistical Mathematics, 44, pp. 97-106.
- [11] McLachlan, G.L. and Peel, D. (2000), *Finite Mixture Models*, New York, Wiley.
- [12] Redner, R.A. and Walker, H.F. (1984) *Mixture Densities, Maximum Likelihood and the EM Algorithm*, SIAM Review, Vol. 26, pp. 195-239.

ON THE BEHAVIOR OF MARGINAL AND CONDITIONAL AKAIKE INFORMATION CRITERIA IN LINEAR MIXED MODELS

Sonja Greven & Thomas Kneib

Johns Hopkins University, Baltimore, USA &

Carl-von-Ossietzky-Universität Oldenburg, Oldenburg, Germany

Abstract: In linear mixed models, the Akaike information criterion (AIC) is often used to decide on the inclusion of a random effect. An important special case is the choice between linear and nonparametric regression models estimated using mixed model penalized splines. We investigate the behavior of two commonly used versions of the AIC, derived either from the implied marginal model or the conditional model formulation. We find that the marginal AIC is not asymptotically unbiased for twice the expected relative Kullback-Leibler distance, and favors smaller models without random effects. For the conditional AIC, it is computationally costly for large sample sizes to correct for estimation uncertainty. However, ignoring it, as is common practice, induces a bias that yields the following behavior: Whenever the random effects variance estimate is positive (even if small), the more complex model is preferred. We illustrate our results in simulations, and investigate their impact in modeling childhood malnutrition in Zambia.

Key Words : Kullback-Leibler information; model selection; penalized splines; random effect; variance component.

1 Introduction

Using penalized splines, linear mixed models can combine model components such as non-linear or spatial effects, surfaces or varying coefficients with cluster-specific random effects. This flexibility in modeling makes model selection increasingly important.

The Akaike information criterion (Akaike, 1973) is often used to decide on inclusion of random effects in mixed models. A common special case is the decision between a linear and a nonparametric function for a covariate effect. An AIC based on the implied marginal likelihood is typically used (mAIC). Vaida & Blanchard (2005) proposed an AIC derived from the conditional model formulation (cAIC). We investigate the behavior of both for the selection of random effects in the linear mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\beta}$ is fixed, \mathbf{b} and $\boldsymbol{\varepsilon}$ are independent, $\mathbf{b} \sim N(\mathbf{0}, \mathbf{D})$ and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

2 The marginal AIC

The AIC can be generally defined as

$$AIC = -2 \log f(\mathbf{y}|\hat{\boldsymbol{\psi}}(\mathbf{y})) + 2E_{\mathbf{y}}[\log f(\mathbf{y}|\hat{\boldsymbol{\psi}}(\mathbf{y})) - \log f(\mathbf{y}|\boldsymbol{\psi}_{\mathbf{K}})] \quad (2)$$

$$+ 2E_{\mathbf{y}}[E_{\mathbf{z}}[\log f(\mathbf{z}|\boldsymbol{\psi}_{\mathbf{K}}) - \log f(\mathbf{z}|\hat{\boldsymbol{\psi}}(\mathbf{y}))]],$$

where $f(\mathbf{y}|\hat{\boldsymbol{\psi}}(\mathbf{y}))$ is the maximized likelihood, and $\boldsymbol{\psi}$ are k unknown parameters with true values $\boldsymbol{\psi}_{\mathbf{K}}$. It is unbiased for twice the expected relative Kullback-Leibler distance. Minimizing (2) thus can be seen as minimizing the average distance of an approximating model to the underlying truth. In standard cases, observations are independent and identically distributed, the parameter space (up to a change of coordinates) is R^k , and the last two terms thus reduce to $2k$ asymptotically. This is the AIC commonly used.

The marginal AIC (mAIC) in the linear mixed model uses the likelihood of the implied marginal model $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ with $\mathbf{V} = \mathbf{I}_n + \mathbf{Z}\mathbf{D}\mathbf{Z}'$. The number of estimable parameters then is $p + q$, with $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ and q the number of unknown parameters $\boldsymbol{\theta}$ in \mathbf{V} . Thus, the mAIC is defined as

$$mAIC = -2 \log(f(\mathbf{y}|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})) + 2(p + q).$$

Now, we can show that due to the marginal correlation structure in \mathbf{y} in (1) and the constraints on $\boldsymbol{\theta}$ (e.g. variances have to be non-negative), the last two terms in (2) are smaller than $2(p + q)$ as well as not independent of the true values in $\boldsymbol{\theta}$. Consequently, the mAIC is positively biased, and favors smaller models without random effects.

3 The conditional AIC

Vaida & Blanchard (2005) define the conditional AIC (cAIC) as

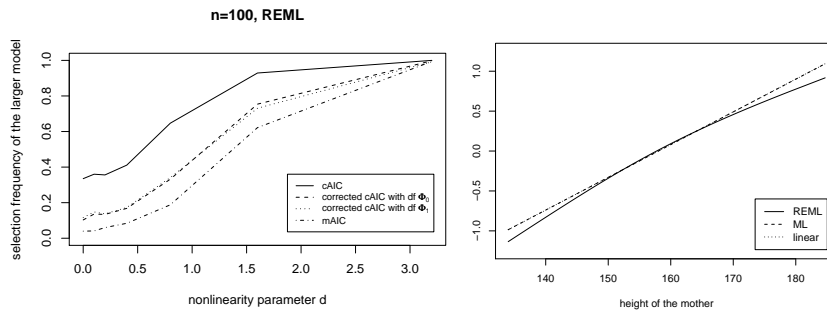
$$cAIC = -2 \log(f(\mathbf{y}|\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}, \hat{\boldsymbol{\theta}})) + 2(\rho + 1),$$

where $f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\theta})$ is the conditional likelihood, $\hat{\mathbf{b}}$ is the BLUP of \mathbf{b} , and

$$\rho = \text{trace} \left(\left(\begin{array}{cc} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{D}_*^{-1} \end{array} \right)^{-1} \left(\begin{array}{cc} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} \end{array} \right) \right).$$

They assume $\mathbf{D}_* = \sigma^{-2}\mathbf{D}$ to be known, but suggest using $\widehat{\mathbf{D}}_*$ otherwise, arguing that the difference is negligible for large n . We call this the simplified cAIC. Liang et al. (2008) propose a corrected cAIC, accounting for estimation of \mathbf{D}_* . For known σ^2 , they replace ρ by $\Phi_0 = \text{trace}(\partial \hat{\mathbf{y}}/\mathbf{y})$, where $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}\hat{\mathbf{b}}$. For unknown σ^2 , Φ_1 involves even second derivatives. As the derivatives are not available in closed form, numerical approximations using n (resp. $2n$) additional model fits have to be used. This can be prohibitive in large samples. In our application, computation time would be about 110 days. As the authors

Figure 1: Selection frequencies of the larger, non-linear model in our simulations for function $m_1(\cdot)$ (left) and estimated effect of height of the mother on the Z-score measuring chronic undernutrition in Zambia (right).



in their simulations find only small differences between simplified and corrected cAIC, we investigate if the often used simplified cAIC is a computationally feasible alternative.

We find the following interesting behavior (for simplicity, we focus on the case of one unknown variance component, i.e. $\mathbf{D} = \tau^2 \mathbf{\Sigma}$): When $\hat{\tau}^2 = 0$, the cAICs of the models including and excluding \mathbf{b} agree, i.e. there is a tie. When $\hat{\tau}^2 > 0$, the cAIC prefers the larger model including \mathbf{b} , regardless of the size of $\hat{\tau}^2$. The simplified cAIC thus is not a useful decision rule, as it does not give guidance on when an estimated variance is large enough to warrant inclusion of the random effect in the model.

4 Simulations

First, we compare a linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with the nonparametric regression model $y_i = m(x_i) + \varepsilon_i$ using mixed model penalized splines. Thus, we select the random effect modeling non-linearity of $m(\cdot)$. The true function is $m_1(x) = 1 + x + 2d(0.3 - x)^2$, $m_2(x) = 1 + x + d(\log(0.1 + 5x) - x)$, or $m_3(x) = 1 + x + 0.3d(\cos(\pi/2 + 2\pi x) - 2x)$ with varying n and non-linearity parameter d . Second, we compare a random intercept and a common intercept model, varying random intercept variance, number and size of clusters.

The simplified cAIC gives a much larger proportion of decisions for the larger model than the mAIC, with the corrected cAIC in-between (Figure 1). While the AIC for nested models in standard settings corresponds to a likelihood ratio test with asymptotic level $\alpha = 0.157$, α is much smaller for the mAIC (as low as 0.01 here), much larger for the simplified cAIC (up to 0.49), and more similar for the corrected cAIC (0.07 to 0.40).

The simplified cAIC chooses the larger model when $\hat{\tau}^2 > 0$, and gives a tie when $\hat{\tau}^2 = 0$. The corrected cAIC often favors the more complex model even when $\hat{\tau}^2 = 0$ due to numerical problems. Especially for Φ_1 using second derivatives, the numerical approximation fails in some cases, resulting in spurious estimated degrees of freedom. Overall, $\Phi_0 + 1$ approximates Φ_1 rather well, but is numerically much more stable.

5 Childhood malnutrition in Zambia

We investigate implications of our findings for model choice in practice. We are interested in modeling the Z-score, measuring chronic undernutrition (stunting) as insufficient height for age, for 1600 children from the 1992 Zambia Demographic and Health Survey. The available predictors are 1) categorical/binary: child's gender, mother's employment status and education 2) spatial: residential district and 3) continuous: duration of breastfeeding, child's age, mother's age, height and BMI. Due to computational cost of the corrected cAIC for large data, we focus only on the mAIC and the simplified cAIC for selecting a random district intercept, and linear or non-linear effects for the continuous variables.

To illustrate our findings, consider the model with height of the mother as the single predictor. Using maximum likelihood estimation, the estimated effect is linear (Figure 1). This results in a tie for the cAIC, while the mAIC clearly prefers the smaller linear model. Using REML estimation, the estimated effect is slightly non-linear. While the mAIC still prefers the smaller, linear model, the cAIC as expected chooses the larger, non-linear model, despite the estimated non-linearity being quite small.

6 Discussion

We investigated the behavior of mAIC and cAIC for selecting random effects in linear mixed models. This corresponds to interesting model choice questions, including decision on non-linearity of effects, constancy of varying coefficients, or the necessity for a random intercept. We found the mAIC to be biased towards simpler models without random effects. The bias is dependent on the setting and the true value of the random effects variance. For the cAIC, it is essential to correct for estimation uncertainty in the unknown random effects covariance matrix. Ignoring the uncertainty, while common and computationally attractive, leads to selection of the random effect whenever it is not estimated to be exactly zero. More research is needed to obtain numerically feasible and robust versions of the corrected cAIC, and to extend methodology to generalized linear mixed models. This paper is based on the technical report at <http://www.bepress.com/jhubiostat/paper179/>.

References

- Akaike, H.** (1973). Information theory and an extension of the maximum likelihood principle. In: *2nd International Symposium on Information Theory*, 267-281.
- Liang, H., Wu, H. and Zou, G.** (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika*, **95**, 773-778.
- Vaida, F. and Blanchard, S.** (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, **92**, 351-370.

PRIORS FOR SPARSE COVARIANCE MODELS

BY KSHITIJ KHARE[†] AND BALA RAJARATNAM^{*}*Stanford University*

Gaussian covariance graph models encode marginal independence among the components of a multivariate random vector by means of a graph G . These models are distinctly different from the traditional concentration graph models (often also referred to as Gaussian graphical models or covariance selection models), as the zeroes in the parameter are now reflected in the covariance matrix Σ , as compared to the concentration matrix $\Omega = \Sigma^{-1}$. The parameter space of interest for covariance graph models is the cone P_G of positive definite matrices with fixed zeroes corresponding to the missing edges of G . As in Letac and Massam [7] we consider the case when G is decomposable. In this paper we construct on the cone P_G a family of Wishart distributions, that serve a similar purpose in the covariance graph setting, as those constructed by Letac and Massam [7] and Dawid and Lauritzen [2] do in the concentration graph setting. We proceed to undertake a rigorous study of these ‘‘covariance’’ Wishart distributions, and derive several deep and useful properties of this class. First, they form a rich conjugate family of priors with multiple shape parameters for covariance graph models. Second, we show how to sample from these distributions by using a block Gibbs sampling algorithm, and prove convergence of this block Gibbs sampler. Development of this class of distributions enables Bayesian inference, which in turn allows for the estimation of Σ even in the case when the sample size is less than the dimension of the data (i.e., when ‘ $n < p$ ’), otherwise not possible in general in the maximum likelihood framework.

1. Introduction. Due to recent advances in science and information technology, there has been a huge influx of high-dimensional data from various fields such as genomics, environmental sciences, finance and the social sciences. Making sense of all the many complex relationships and multivariate dependencies present in the data and formulating correct models and developing inferential procedures is one of the major challenges in modern day statistics. In parametric models the covariance or correlation matrix (or its inverse) is the fundamental object that quantifies relationships between random variables. Estimating the covariance matrix in a sparse way is crucial in high dimensional problems and enables the detection of the most important relationships. In this light, graphical models have served as tools to discover structure in high-dimensional data.

The primary aim of this note is to develop a new family of conjugate prior distributions for covariance graph models (a subclass of graphical models), and consequently study the properties of this family of distributions. These properties are highly attractive for Bayesian inference in high-dimensional settings. In covariance graph models, specific entries of the covariance matrix are restricted to be zero, which implies marginal independence in the Gaussian case. Covariance graph models correspond to curved exponential families, and are distinctly different from the well-studied concentration graph models, which in turn correspond to natural exponential families. A rich framework for Bayesian inference for natural exponential

^{*}Bala Rajaratnam was supported in part by NSF grant DMS 0505303, DMS 0906392, SUFSC08-SUSHSTF09-SMSCVISG0906

[†]Kshitij Khare was supported in part by the B.C. and E.J. Eaves Stanford graduate fellowship.

families has been established in the last three decades, starting with the seminal and celebrated work of Diaconis and Ylvisaker [3] that laid the foundations for constructing conjugate prior distributions for natural exponential family models. The Diaconis-Ylvisaker (henceforth referred to as DY) conjugate priors are characterized by posterior linearity. An analogous framework for curved exponential families is not available in the literature. This note develops a framework for priors for the class of covariance graph models and summarizes some of the main results from a longer paper (see [5] for more details).

2. Preliminaries. An undirected graph G is a pair (V, E) , where V is a permutation¹ of the set $\{1, 2, \dots, m\}$ denoting the set of vertices of G . The set $E \subseteq V \times V$ denotes the set of edges in the graph. If vertices u and v are such that $(u, v) \in E$ then we say that there is an edge between u and v . It is also understood that $(u, v) \in E$ implies $(v, u) \in E$, i.e., the edges are undirected. Though the dependence of $G = (V, E)$ on the particular ordering in V is often suppressed, the reader should bear in mind that unlike traditional graphs, the graphs defined above are not equivalent up to permutation of the vertices² modulo the edge structure. We describe below two classes of graphs which play a central role in this paper.

2.1. Decomposable graphs. An undirected graph G is said to be *decomposable* if any induced subgraph does not contain a cycle of length greater than or equal to four. The reader is referred to Lauritzen [6] for all the common notions of graphical models (and in particular decomposable graphs) that we will use here. One such important notion is that of a perfect order of the cliques. Every decomposable graph admits a perfect order of its cliques. Let (C_1, C_2, \dots, C_k) be one such perfect order of the cliques of the graph G . The *history* for the graph is given by $H_1 = C_1$ and

$$H_j = C_1 \cup C_2 \cup \dots \cup C_j, \quad j = 2, 3, \dots, k,$$

and the *minimal separators* of the graph are given by

$$S_j = H_{j-1} \cap C_j, \quad j = 2, 3, \dots, k.$$

Let

$$R_j = C_j \setminus H_{j-1} \text{ for } j = 2, 3, \dots, k.$$

Let $k' \leq k - 1$ denote the number of distinct separators and $\nu(S)$ denote the multiplicity of S , i.e., the number of j such that $S_j = S$. Generally, we will denote by \mathcal{C} the set of cliques of a graph and by \mathcal{S} its set of separators.

Now, let Σ be an arbitrary positive definite matrix with zero restrictions according to $G = (V, E)$ ³, i.e., $\Sigma_{ij} = 0$ whenever $(i, j) \notin E$. It is known that if G is decomposable, there exists an ordering of the vertices such that if $\Sigma = LDL^T$ is the modified Cholesky decomposition corresponding to this ordering, then for $i > j$,

$$(2.1) \quad L_{ij} = 0 \text{ whenever } (i, j) \notin E.$$

Lauritzen [6] shows that this ordering provides a ‘perfect vertex elimination scheme’ for G .

¹The ordering in V is emphasized here as the elements of V will later correspond to rows or columns of matrices.

²This has been done for notational convenience, as will be seen later.

³It is emphasized here that the ordering of the vertices reflected in V plays a crucial role in the definitions and results that follow.

2.2. *The spaces P_G , Q_G and \mathcal{L}_G .* An m -dimensional Gaussian covariance graph model⁴ can be represented by the class of multivariate normal distributions with fixed zeros in the covariance parameter (i.e., marginal independencies) described by a given graph $G = (V, E)$. That is, if $(i, j) \notin E$, the i -th and j -th components of the multivariate random vector are marginally independent. Without loss of generality, we can assume that these models have mean zero and are characterized by the parameter set P_G of positive definite covariance matrices Σ such that $\Sigma_{ij} = 0$ whenever the edge (i, j) is not in E . Following the notation in [7, 9] for G decomposable, we define Q_G to be the space on which the free elements of the precision matrices (or inverse covariance matrices) Ω live.

More formally let M denote the set of symmetric matrices of order m , $M_m^+ \subset M$ the cone of positive definite matrices (abbreviated > 0), I_G the linear space of symmetric incomplete matrices x with missing entries $x_{ij}, (i, j) \notin E$ and $\kappa : M \mapsto I_G$ the projection of M into I_G . The parameter set of the precision matrices of Gaussian covariance graph models can also be described as the set of incomplete matrices $\Omega = \kappa(\Sigma^{-1})$, $\Sigma \in P_G$. We are therefore led to consider the two cones

$$(2.2) \quad P_G = \{y \in M_m^+ \mid y_{ij} = 0, (i, j) \notin E\}$$

$$(2.3) \quad Q_G = \{x \in I_G \mid x_{C_i} > 0, i = 1, \dots, k\}.$$

where $P_G \subset Z_G$ and $Q_G \subset I_G$, where Z_G denotes the linear space of symmetric matrices with zero entries $y_{ij}, (i, j) \notin E$.

Gróne *et al.* [4] proved formally that there is an isomorphism⁵ between P_G and Q_G .

We now introduce new spaces that we shall need in our subsequent analysis⁶. Let \mathcal{L}_G denote the space of all lower triangular matrices with diagonal entries equal to 1, such that the entries in the lower triangle have zero restrictions corresponding to G , i.e.,

$$\mathcal{L}_G = \{L : L_{ij} = 0 \text{ whenever } i < j, \text{ or } (i, j) \notin E, \text{ and } L_{ii} = 1, \forall 1 \leq i, j \leq m\}.$$

Define Θ_G (the modified Cholesky space) by

$$\Theta_G = \{\theta = (L, D) : L \in \mathcal{L}_G, D \text{ diagonal with } D_{ii} > 0 \forall 1 \leq i \leq m\}.$$

Denote the mapping

$$\psi : \Theta_G \rightarrow M_m^+$$

as defined by

$$(2.4) \quad \psi(L, D) = LDL^T$$

The above mapping ψ plays an important role in our analysis, and shall be studied later.

⁴A brief overview of the literature in this area is provided in the introduction.

⁵Furthermore, it also defines a diffeomorphism.

⁶These spaces are not defined in [7, 9].

2.3. Vertex ordering. Let $G = (V, E)$ be an undirected decomposable graph with vertex set $V = \{1, 2, \dots, m\}$ and edge set E . Let S_V denote the permutation group associated with V . For any $\sigma \in S_V$, let $G_\sigma := (\sigma(V), E_\sigma)$, where $(u, v) \in E_\sigma$ iff $(\sigma^{-1}(u), \sigma^{-1}(v)) \in E$. Let $S_D \subset S_V$ denote the subset of permutations σ of V , such that for any $\Sigma \in M_m^+$ with $\Sigma = LDL^T$, $L \in \mathcal{L}_{G_\sigma} \Leftrightarrow \Sigma \in P_{G_\sigma}$. Hence for every $\sigma \in S_D$, the mapping $\psi_\sigma : \Theta_{G_\sigma} \rightarrow M_m^+$ defined in (2.4), is a bijection from Θ_{G_σ} to P_{G_σ} . In particular, the ordering corresponding to any perfect vertex elimination scheme lies in S_D . If G is homogeneous, let $S_H \subset S_D$ denote the subset of permutations σ of V , such that $L \in \mathcal{L}_{G_\sigma} \Leftrightarrow L^{-1} \in \mathcal{L}_{G_\sigma}$. In particular, any ordering of the vertices corresponding to the Hasse perfect vertex elimination scheme lies in S_H . The above defines a nested triplet of permutations of V given by $S_H \subset S_D \subset S_V$.

3. Wishart distributions for covariance graphs. Let $G = (V, E)$ be an undirected decomposable graph with vertex set V and edge set E . We assume that the vertices in V are ordered so that $V \in S_D$. The covariance graph model associated with G is the family of distributions

$$\begin{aligned} \mathcal{G} &= \{\mathcal{N}_m(\mathbf{0}, \Sigma) : \Sigma \in P_G\} \\ &\cong \{\mathcal{N}_m(\mathbf{0}, LDL^T) : (L, D) \in \Theta_G\}. \end{aligned}$$

Consider the class of measures on Θ_G with density (w.r.t. $\prod_{i>j, (i,j) \in E} dL_{ij} \prod_{i=1}^m dD_{ii}$)

$$(3.1) \quad \tilde{\pi}_{U, \alpha}(L, D) = e^{-\frac{(\text{tr}((LDL^T)^{-1}U) + \sum_{i=1}^m \alpha_i \log D_{ii})}{2}}, \quad \theta = (L, D) \in \Theta_G.$$

These measures are parameterized by a positive definite matrix U and a vector $\alpha \in \mathbb{R}^m$ with non-negative entries. Let us first establish some notation.

- $\mathcal{N}(i) := \{j : (i, j) \in E\}$
- $\mathcal{N}^{\prec}(i) := \{j : (i, j) \in E, i > j\}$
- $U^{\prec i} := ((U_{kl}))_{k, l \in \mathcal{N}^{\prec}(i)}$
- $U^{\preceq i} := ((U_{kl}))_{k, l \in \mathcal{N}^{\prec}(i) \cup \{i\}}$
- $U_{\cdot i}^{\prec} := (U_{ki})_{k \in \mathcal{N}^{\prec}(i)}$

Let

$$z_G(U, \alpha) := \int e^{-\frac{(\text{tr}((LDL^T)^{-1}U) + \sum_{i=1}^m \alpha_i \log D_{ii})}{2}} dL dD.$$

If $z_G(U, \alpha) < \infty$, then $\tilde{\pi}_{U, \alpha}$ can be normalized to obtain a probability measure. A sufficient condition for the existence of a normalizing constant for $\tilde{\pi}_{U, \alpha}(L, D)$ is provided in the following proposition.

THEOREM 1. *Let $dL := \prod_{(i,j) \in E, i>j} dL_{ij}$ and $dD := \prod_{i=1}^m dD_{ii}$. Then,*

$$\int_{\Theta_G} e^{-\frac{(\text{tr}((LDL^T)^{-1}U) + \sum_{i=1}^m \alpha_i \log D_{ii})}{2}} dL dD < \infty$$

if

$$\alpha_i > |\mathcal{N}^{\prec}(i)| + 2 \quad \forall i = 1, 2, \dots, m.$$

LEMMA 1. Let $G = (V, E)$ be a decomposable graph, where vertices in V are ordered so that $V \in S_D$. Let $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$ be an i.i.d. sample from $\mathcal{N}_m(\mathbf{0}, LDL^T)$, where $(L, D) \in \Theta_G$. Let $S = \frac{1}{n} \sum_{i=1}^n \mathbf{Y}_i \mathbf{Y}_i^T$ denote the empirical covariance matrix. If the prior distribution on (L, D) is $\pi_{U, \alpha}$, then the posterior distribution of (L, D) is given by $\pi_{\tilde{U}, \tilde{\alpha}}$, where $\tilde{U} = nS + U$ and $\tilde{\alpha} = (n + \alpha_1, n + \alpha_2, \dots, n + \alpha_m)$.

Proof. The likelihood of the data is given by

$$f(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n \mid L, D) = \frac{1}{(\sqrt{2\pi})^{nm}} e^{-\frac{\text{tr}((LDL^T)^{-1}(nS)) + n \log |D|}{2}}.$$

Using $\pi_{U, \alpha}$ as a prior for (L, D) , the posterior distribution of (L, D) given the data $(\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n)$ is

$$\pi_{U, \alpha}(L, D \mid \mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n) \propto e^{-\frac{\text{tr}((LDL^T)^{-1}(nS+U)) + \sum_{i=1}^m (n+\alpha_i) \log D_{ii}}{2}}, \quad \theta \in \Theta_G.$$

Hence the posterior distribution belongs to the same family as the prior, i.e.,

$$\pi_{U, \alpha}(\cdot \mid \mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n) = \pi_{\tilde{U}, \tilde{\alpha}}(\cdot),$$

where $\tilde{U} = nS + U$ and $\tilde{\alpha} = (n + \alpha_1, n + \alpha_2, \dots, n + \alpha_m)$. \square

Remark. If we assume that the observations do not have mean zero, i.e., $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$ are i.i.d. $\mathcal{N}(\mu, \Sigma)$, with $\mu \in \mathbb{R}^m$, $\Sigma \in P_G$, then

$$\tilde{S} := \frac{1}{n} \sum_{i=1}^n (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^T$$

is the minimal sufficient statistic for Σ . Here, $n\tilde{S}$ has a Wishart distribution with parameter Σ and $n - 1$ degrees of freedom. Hence, if we assume a prior $\pi_{U, \alpha}$ for (L, D) , then the posterior distribution is given by

$$\pi_{U, \alpha}(\cdot \mid \tilde{S}) = \pi_{\tilde{U}, \tilde{\alpha}}(\cdot),$$

where $\tilde{U} = n\tilde{S} + U$, and $\tilde{\alpha} = (n - 1 + \alpha_1, n - 1 + \alpha_2, \dots, n - 1 + \alpha_m)$.

3.1. *Induced prior on P_G and Q_G .* The prior $\pi_{U, \alpha}$ on Θ_G (the modified Cholesky space) induces a prior on P_G (the covariance matrix space) and Q_G . Recall from Section 2.2 that P_G is the space of positive definite matrices with zero restrictions according to G , and Q_G is the space of incomplete matrices isomorphic to P_G . We provide an expression for the induced priors on these spaces in order to compare our Wishart distributions with other classes of distributions. Note that since the vertices have been ordered so that $V \in S_D$, the transformation

$$\psi : \Theta_G \rightarrow M_m^+$$

defined by

$$\psi(L, D) = LDL^T =: \Sigma$$

is a bijection from Θ_G to P_G . The lemma below provides the required Jacobians for deriving the induced priors on P_G and Q_G . We prove the first part, and the proof of the second part

can be found in [10]. The reader is referred to Section [9] for notation on decomposable graphs. Note that if x is a matrix, $|x|$ denotes its determinant, while if C is a set, then $|C|$ denotes its cardinality.

LEMMA 2. *Jacobians of transformations.*

1. The Jacobian of the transformation $\psi : (L, D) \rightarrow \Sigma$ from Θ_G to P_G is

$$\prod_{i=1}^m D_{jj}(\Sigma)^{-n_j}.$$

Here $D_{jj}(\Sigma)$ denotes that D_{jj} is a function of Σ , and $n_j := |\{i : (i, j) \in E, i > j\}|$ for $j = 1, 2, \dots, m$.

2. The absolute value of the Jacobian of the bijection $\zeta : x \rightarrow \hat{x}^{-1}$ from Q_G to P_G is

$$\prod_{C \in \mathcal{C}} |x_C|^{-|C|-1} \prod_{S \in \mathcal{S}} |x_S|^{(|S|+1)\nu(S)}.$$

These Jacobians allow us to compute the induced priors on P_G and Q_G . The induced prior corresponding to $\tilde{\pi}_{U,\alpha}$ on P_G is given by

$$(3.2) \quad \tilde{\pi}_{U,\alpha}^{P_G}(\Sigma) \propto e^{-\frac{(\text{tr}(\Sigma^{-1}U) + \sum_{i=1}^m (2n_i + \alpha_i) \log D_{ii}(\Sigma))}{2}}, \quad \Sigma \in P_G.$$

We first note that the traditional Inverse-Wishart distribution (see [8]) with parameters U and n is a special case of (3.2) when G is the complete graph, and $\alpha_i = n - 2m + 2i$, $\forall 1 \leq i \leq m$. We also note that the \mathcal{G} -inverse Wishart priors introduced in [11] have a one-dimensional shape parameter δ , and are a very special case of our richer class $\tilde{\pi}_{U,\alpha}^{P_G}$. The single shape parameter δ is given by the relationship $\alpha_i + 2n_i = \delta + 2m$, $1 \leq i \leq m$ ⁷.

We now proceed to derive the induced prior on Q_G . Let $x = \varphi(\Sigma) = \kappa(\Sigma^{-1})$ denote the image of Σ in Q_G . Using the second part of Lemma 2, the induced prior corresponding to $\tilde{\pi}_{U,\alpha}$ on Q_G is given by

$$\begin{aligned} \tilde{\pi}_{U,\alpha}^{Q_G}(x) &\propto e^{-\frac{(\text{tr}(\hat{x}U) + \sum_{i=1}^m (2n_i + \alpha_i) \log D_{ii}(\hat{x})^{-1})}{2}} \\ &\times \frac{\prod_{S \in \mathcal{S}} |x_S|^{(|S|+1)\nu(S)}}{\prod_{C \in \mathcal{C}} |x_C|^{|C|+1}}, \quad x \in Q_G. \end{aligned}$$

4. Sampling from the posterior distribution. In this section, we study the properties of our family of distributions, and thereby provide a method that allows us to generate samples from the posterior distribution corresponding to the priors defined in Section 3. In

⁷There is an interesting parallel here that becomes apparent from our derivations above. In the concentration graph setting, the single shape parameter hyper inverse Wishart (HIW) prior of Dawid and Lauritzen [2] is a special case of the multiple shape parameter class of priors introduced by Letac and Massam [7], in the sense $\alpha_i = -\frac{1}{2}(\delta + c_i - 1)$ (see [9] for notation). In a similar spirit, we discover that the single shape parameter class of priors in [11] is a special case of the multiple shape parameter class of priors $\tilde{\pi}_{U,\alpha}$ introduced in this paper, in the sense $\alpha_i = \delta - 2n_i + 2m$.

particular, we prove that $\theta = (L, D) \in \Theta_G$ can be partitioned into blocks so that the conditional distribution of each block given the others are standard distributions in statistics and hence easy to sample from. We can therefore generate samples from the posterior distribution by using the block Gibbs sampling algorithm.

4.1. *Distributional properties and the block Gibbs sampler.* Let us introduce some notation before deriving the required conditional distributions. Let $G = (V, E)$ be a decomposable graph, such that $V \in S_D$. For a lower triangular matrix L with diagonal entries equal to 1,

$$\begin{aligned} L_{u\cdot} &:= u^{\text{th}} \text{ row of } L, \quad u = 1, 2, \dots, m, \\ L_{\cdot v} &:= v^{\text{th}} \text{ column of } L, \quad v = 1, 2, \dots, m, \\ L_{\cdot v}^G &:= (L_{uv})_{u>v, (u,v) \in E}, \quad v = 1, 2, \dots, m-1. \end{aligned}$$

So $L_{\cdot v}^G$ is the v^{th} column of L without the components which are specified to be zero under the model \mathcal{G} (and without the v^{th} diagonal entry, which is 1). In terms of this notation, the parameter space can be represented as

$$(4.1) \quad \Theta_G = \left\{ (L_{\cdot 1}^G, L_{\cdot 2}^G, L_{\cdot 3}^G, \dots, L_{\cdot m-1}^G, D) : L_{ij} \in \mathbb{R}, \forall 1 \leq j < i \leq m, (i, j) \in E, D_{ii} > 0, \forall 1 \leq i \leq m \right\}.$$

Suppose $\theta \sim \pi_{U, \alpha}$ for some positive definite U and $\alpha \in \mathbb{R}^m$ with non-negative entries. Then the posterior distribution is $\pi_{\tilde{U}, \tilde{\alpha}}$, where $\tilde{U} = nS + U, \tilde{\alpha} = (n + \alpha_1, n + \alpha_2, \dots, n + \alpha_m)$. In the following proposition, we derive the distributional properties which provide the essential ingredients for our block Gibbs sampling procedure.

THEOREM 2. *Using the notation above, the conditional distributions of each component of θ (as in (4.1)) given the other components and the data $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$ are as follows.*

1.

$$L_{\cdot v}^G \mid (L \setminus L_{\cdot v}^G, D, \mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n) \sim \mathcal{N}(\mu^{v,G}, M^{v,G}) \quad \forall v = 1, 2, \dots, m-1,$$

where

$$\mu_u^{v,G} := \mu_u^v + \sum_{u' > v: (u', v) \in E} \sum_{\substack{w > v: (w, v) \notin E \\ \text{OR } w < v, L_{vw}^{-1} = 0}} M_{uu'}^{v,G} \left(L^{-1} \tilde{U} (L^T)^{-1} \right)_{vv} (LDL^T)_{u'w}^{-1} \mu_w^v$$

$$\forall u > v, (u, v) \in E,$$

$$\mu_u^v := \frac{(L^{-1} \tilde{U})_{vu}}{\left(L^{-1} \tilde{U} (L^T)^{-1} \right)_{vv}} \quad \forall u \text{ such that } L_{vu}^{-1} = 0,$$

$$(M^{v,G})_{uu'}^{-1} := \left(L^{-1} \tilde{U} (L^T)^{-1} \right)_{vv}^{-1} (LDL^T)_{uu'}^{-1} \quad \forall u, u' > v, (u, v), (u', v) \in E.$$

2.

$$D_{ii} \mid L, \mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_m \sim IG \left(\frac{\tilde{\alpha}_i}{2} - 1, \frac{\left(L^{-1} \tilde{U} (L^T)^{-1} \right)_{ii}}{2} \right)$$

independently for $i = 1, 2, \dots, m$, where IG represents the inverse-gamma distribution.

Remark The notation $L_{vw}^{-1} = 0$ in the definition of $\mu^{v,G}$ above means indices w for which L_{vw}^{-1} is 0 as a function of entries of L .

4.2. Convergence of block Gibbs sampler.

We now prove that sufficient conditions for convergence of a Gibbs sampling Markov chain to its stationary distribution (see [1, Theorem 6]) are satisfied by the Markov chain corresponding to our block Gibbs sampler. Let $\phi(\mathbf{x} \mid \mu, \Sigma)$ denote the $\mathcal{N}(\mu, \Sigma)$ density evaluated at \mathbf{x} . Let $f_{IG}(d \mid \alpha, \lambda)$ denote the $IG(\alpha, \lambda)$ density evaluated at d . Let us fix $\psi, d_1, d_2 > 0$ arbitrarily. Let

$$\Theta_{\psi, d_1, d_2} := \{\theta = (L, D) \in \Theta_G : |L_{ij}| \leq \psi, d_1 \leq D_{ii} \leq d_2 \forall i > j, (i, j) \in E\}.$$

We now formally prove the conditions which are sufficient for establishing convergence.

PROPOSITION 1. $\exists \delta > 0$ such that for all $\theta = (L, D) \in \Theta_{\psi, d_1, d_2}$,

$$\begin{aligned} \phi\left(L_{:,v}^G \mid \mu^{v,G}, M^{v,G}\right) &> \delta \forall v = 1, 2, \dots, m-1, \\ f_{IG}\left(D_{ii} \mid \frac{\tilde{\alpha}_i}{2} - 1, \frac{\left(L^{-1}\tilde{U}(L^T)^{-1}\right)_{ii}}{2}\right) &> \delta \forall i = 1, 2, \dots, m. \end{aligned}$$

The reader is referred to a longer paper (see [5]) where proofs of the above theorems/propositions are provided.

References.

- [1] Athreya, K.B., Doss, H. and Sethuraman, J. (1996). On the convergence of the Markov chain simulation method, *Ann. Statist.* **24**, 69-100.
- [2] Dawid, A.P. and Lauritzen, S.L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models, *Ann. Statist.* **21**, 1272-1317.
- [3] Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families, *Ann. Statist.* **7**, 269-281.
- [4] Grone, R., Johnson, C.R., S'a, E.M. and Wolkowicz, H. (1984). Positive definite completions of partial hermitian matrices, *Linear Algebra Appl.* **58**, 109124.
- [5] Khare, K. and Rajaratnam, B. (2008). Wishart distributions for covariance graph models. Technical Report No. 2008-11, Department of Statistics, Stanford University.
- [6] Lauritzen, S.L. (1996). *Graphical models*, Oxford University Press Inc., New York.
- [7] Letac, G. and Massam, H. (2007). Wishart distributions for decomposable graphs, *Ann. Statist.* **35**, 1278-1323.
- [8] Muirhead, R.J. (1982). *Aspects of multivariate statistical theory*, John Wiley and Sons, New York.
- [9] Rajaratnam, B., Massam, H. and Carvalho, C. (2008). Flexible covariance estimation in graphical models, *Annals of Statistics* **36**, 2818-2849.
- [10] Roverato, A. (2000). Cholesky decomposition of a hyper inverse Wishart matrix, *Biometrika* **87**, 99-112.
- [11] Silva, R. and Ghahramani, Z. (2008). The hidden life of latent variables: Bayesian learning with mixed graph models. Preprint, Cambridge University.

DEPARTMENT OF STATISTICS
STANFORD, CA 94305-4065, U.S.A.
E-MAIL: kdkhare@stanford.edu

DEPARTMENT OF STATISTICS
STANFORD, CA 94305-4065, U.S.A.
E-MAIL: brajarat@stanford.edu

EXPLORATORY ANALYSIS OF MULTIVARIATE DEPENDENCIES IN HIGH DIMENSIONAL DATA

Kshitij Khare, Bala Rajaratnam & Julia Salzman

*Stanford University, California, USA & Stanford University School of Medicine,
California, USA*

Abstract : In recent years, the availability of high-throughput data from genomic, finance, environmental, and other applications has created an urgent need for methodology and tools for analyzing high-dimensional data. Making sense of the many complex relationships, formulating correct models and developing inferential procedures is one of the major challenges facing statisticians today, and also those working in applied fields. The co-variance parameter (or its inverse) is a natural parameter of interest when trying to understand complex relationships between many variables. We propose a new method to estimate the inverse covariance matrices in a sparse manner when there is a natural order in the data. We compare our method to others in the literature to assess its effectiveness in high dimensional problems.

Key Words : Exploratory Data Analysis, High Dimensional Data, Gaussian Model

1 Introduction

A difficult and important problem in genetics is to estimate clusters of dependent mRNA expression levels among the many thousands of expressed transcripts. The approach taken here is to introduce and analyze several exploratory techniques which attempt to recover the inverse covariance matrix and hence find sets of genes for which there is evidence of conditional independence. It is particular interest to estimate the zeros of the inverse covariance matrix as zeros correspond to lack of edges and hence conditional independence relationships. The work here is motivated by the Generalized Topological Overlap (GTOM) measure developed by Horvath et al, a measure which attempts to quantify multivariate dependency without directly estimating a concentration graph.

Two applications for exploratory estimates of the concentration graph motivating the development of GTOM and the measures introduced here are 1) variable clustering and 2) estimates of the concentration graph, perhaps marginal estimates of each cluster found in 1). Variable clustering, or identifying groups of variables which are highly conditionally dependent, has the potential to identify significant co-regulated genes.

We have developed an alternative exploratory class of estimators of the inverse covariance matrix (denoted POM, IPOM) which can be used for exploratory analysis to both estimate the concentration graph in Gaussian graphical models and for clustering variables (genes). These estimators bypass the difficult problem of model selection for graphical models. Some theoretical properties are presented in this proceedings. Applications to variable clustering are committed due to space limitations. A special case of POM and IPOM coincide with an algorithmic approach to inverse covariance matrix estimation proposed by Friedman et al (unpublished).

2 Methodology: Definitions of POM, IPOM

Definition 2.1. Let $P := I - S$, then

$$IPOM_k = I + P - P^2 + \dots + (-1)^k P^k.$$

Definition 2.2. Let $P := I - S$, then

$$POM_k = I + P + P^2 + \dots + P^k.$$

Note that:

Remark 2.3. For $i \neq j$,

$$IPOM_1(S)_{ij} = POM_1(S)_{ij} = -S_{ij}.$$

2.1 Comparison and Interpretation of POM, IPOM, GTOM

Despite potential empirical success of GTOM, the following reasons suggest a refinement of methodology is useful:

- GTOM is a function of a binary adjacency matrix whereas IPOM, POM are functions of any estimate of the covariance matrix.
- GTOM is designed for clustering and not estimation of the inverse covariance matrix; fails for estimating the inverse covariance matrix in a simple cases of separator of size 1, or for a two variable $AR(2)$ process.

The two estimators POM_k and $IPOM_K$ are extensions of GTOM in the following manner:

- POM is a generalization of the numerator of GTOM both of which attempt to estimate clusters in the covariance matrix.

- The numerator, denominator of GTOM, and hence the estimator GTOM may be unstable since they are functions of a threshold function on a sample covariance graph.
- IPOM can be used for a more direct approach of estimating graph structure and then doing graph based clustering.

3 Theoretical Analysis for GTOM, IPOM

There is current interest in so called “hub genes” in both biology and other fields. One motivation for the use of $GTOM_i$ is its use in estimating “hub genes”. However, this section demonstrates theoretical reasons that $GTOM$ is unable to detect hub genes for $m = 1$.

Recall the definition of $GTOM_1$:

Definition 3.1. For $N_m(i) = \{j \neq i | \min \text{ path length}(i,j) \leq m\}$,

$$GTOM_m(i, j) = \frac{|N_1(i) \cap N_1(j)| + a_{ij}}{\min(N_1(i), N_1(j)) + 1 - a_{ij}}$$

Consider A whose adjacency matrix is a star graph with n vertices and hub labeled as 1. For simplicity, assume all edge weights on the inverse covariance matrix are uniform. Such an adjacency graph arises as both the concentration graph of a Gaussian graphical model with a “hub” and as the thresholded version of the corresponding covariance matrix.

Claim 3.2 (Star Graph). For A the star graph on n vertices with hub vertex 1, for $m = 1, GTOM_m(i, j) = 1$ for all i, j . Otherwise, $GTOM_m(i, j) = 1$ if $i = 1$ or $j = 1$ and otherwise $GTOM_m(i, j) = \frac{n-2}{n}$.

Proof of Claim 3.2. The case for $m = 1$ is straightforward. There are only 2 cases to check for $i < j$. Case 1: $i = 1$. In this case, The denominator of $GTOM_1$ is 1 and the numerator is also 1. In the case $i > 1$, the numerator of $GTOM_1$ is 1 as is the denominator. For $m > 1$, $N_m(i) = \{n/i\}$ for all i, m . The computation is straightforward. \square

The above argument shows that as n becomes large, $GTOM_m(i, j)$ approaches 1 for all i, j . The $IPOM_1$ and POM_1 measures each produce an estimate of the covariance matrix which differentiates the “hub” node from other nodes.

The following Propositions provide a theoretical foundation for the estimator $IPOM_k$. The first establishes IPOM as a plug-in estimator for a class of covariance matrices specified in the

Proposition 3.3. Suppose Σ is an p by p covariance matrix with $\|\Sigma\| < 2$ and $\Sigma_{ii} = 1$ for all $1 \leq i \leq p$. Then, $IPOM_k$ is the plug-in estimator for the k^{th} order Taylor series approximation of the inverse, Σ_k^{-1} , which has the property that

$$\|\Sigma^{-1} - \Sigma_k^{-1}\|_\infty < \frac{(\lambda_{max} - 1)^k}{1 - \lambda_{max}}.$$

where $\lambda_{max} = \|\Sigma^{-1}\| - 1$.

The next Propositions establish probabilistic properties of a thresholded version of IPOM made precise in the following

Definition 3.4. Define $TIPOM_1(S, \lambda)$ as $|IPOM_1(S)|$ which has been thresholded at a value $\lambda < 1$.

The next proposition shows that $TIPOM_1(S)$ will recover the inverse covariance matrix of Gaussian graphical models with maximum clique size 2.

Proposition 3.5. For Σ^{-1} a decomposable connected graph of maximum clique size 2, for $0 < a < b < 1$, if

$$\sigma_{i,j} \in \{0, [a, b], [-b, -a]\} \quad \text{whenever } (i, j) \in \mathcal{E} \quad (1)$$

where $b^2 < a - \epsilon$ for some $\epsilon > 0$ and $\sigma_{ii} = 1$ for all i , then there exists a $\lambda \in (0, 1)$ so that asymptotically (and with a probability we can bound in finite samples), $TIPOM_1(S, \lambda)$ recovers the zero structure of Σ^{-1} . Without the assumption of Equation 1, the conclusion does not necessarily hold.

The final Proposition shows that if the inverse covariance matrix is the sum of a matrix corresponding to a Gaussian graphical model with maximum clique size 2 and an error term distributed, up to determined matrix multiplication, according to the GOE, an analogue of Proposition 3.5 holds.

Suppose $\Omega_0 = \Sigma_0^{-1}$ is the inverse covariance matrix for a graph whose maximum clique size is two and $\epsilon > 0$ is fixed and to be determined later. Let

$$\Sigma^{-1} = \Omega_0^{\frac{1}{2}}(I + \epsilon O^2)\Omega_0^{\frac{1}{2}} \quad (2)$$

where O is a random matrix with distribution of the GOE. Note that O^2 is symmetric and positive definite. Therefore,

Proposition 3.6. Let Ω_0 be the inverse covariance of a Gaussian graphical model with maximum clique size 2, minimum eigenvalue $\frac{1}{\gamma^2}$. Suppose $\Sigma_0 = \Omega_0^{-1}$ a matrix whose entries are in the set $\{[a, b], 0, [-b, -a]\}$ where $b^2 + x < a - x$ for some $x > 0$, $\epsilon^2(4 + y) < 1$ and

$$x > \frac{1}{\gamma_1^2} \frac{\epsilon^2(4 + y)}{1 - \epsilon^2(4 + y)}.$$

Then, there is a λ so that the procedure $TIPOM(\cdot, \lambda)$ will recover the edges in Ω_0 .

A Note on Congruence Relations for Second-Order Processes

David King

*School of Mathematics and Statistics, Arizona State University, Tempe, AZ
85287-1804*

Abstract

Explicit formulas are derived for the congruence mappings that connect three Hilbert spaces associated with a second-order stochastic process. In particular, an insightful expression is obtained for the mapping that connects a process to its corresponding reproducing kernel Hilbert space.

Key words: covariance operator, H-valued random variables, reproducing kernel Hilbert space

AMS 2000 Subject Classification: Primary 62H20, 62H25, 62M99

1 Introduction

Let $\{X(t), t \in E\}$ be a zero mean stochastic process with a finite dimensional index set $E = \{t_1, \dots, t_n\}$. The set of all linear combinations of $\mathbf{X} = (X(t_1), \dots, X(t_n))'$ is isometrically isomorphic or congruent to the column space of the covariance matrix $\mathbf{K} = \{K(t_i, t_j)\}_{i,j=1}^n$, where K is the covariance kernel defined by

$$K(s, t) = E[X(s)X(t)].$$

The congruence mapping is determined uniquely by $\Psi(K(t, \cdot)) = X(t), t \in E$, with the result that every linear combination U of the \mathbf{X} vector with nonzero variance can be expressed as

$$U = \Psi(\mathbf{f}) = \mathbf{f}'\mathbf{K}^\dagger\mathbf{X} \tag{1}$$

Email address: dbking@asu.edu (David King).

for some $\mathbf{f} \in \ker(\mathbf{K})^\perp$, with \mathbf{K}^\dagger denoting the Moore–Penrose generalized inverse of \mathbf{K} .

The congruence in (1) is a special case of the Løve–Parzen congruence that connects a second order process to the reproducing kernel Hilbert space (RKHS) generated by its covariance kernel. In general, however, there is no simple closed form for the congruence mapping. This poses problems for the application of RKHS methods in areas such as functional data analysis (FDA) (e.g., Eubank and Hsing 2008). Here we deal with a case of particular interest for FDA settings and show that (1) has a natural infinite dimensional extension for processes that take values in certain Hilbert function spaces.

Let E be a subset of \mathbb{R} and ν a sigma-finite measure on E . We then consider the case where a zero-mean stochastic process $\{X(t), t \in E\}$ takes values in the Hilbert space $\mathcal{H} = L^2(E)$ of square integrable functions on E with inner product $(f, g)_\mathcal{H} \equiv \int_E f(t)g(t)d\nu(t)$. Associated with $X(\cdot)$ is the covariance operator $S : \mathcal{H} \mapsto \mathcal{H}$ defined by

$$(f, Sg)_\mathcal{H} \equiv \text{Cov}((X, f)_\mathcal{H}, (X, g)_\mathcal{H}) = \int_{\mathcal{H}} (X, f)_\mathcal{H}(X, g)_\mathcal{H} d\mathcal{P}(X) \quad \text{for } f, g \in \mathcal{H},$$

with \mathcal{P} the induced probability measure from X on \mathcal{H} . It is well known (e.g. Laha and Rohatgi 1979) that S is positive, self-adjoint and Hilbert–Schmidt. As a result, X admits a Karhunen–Loève expansion $X(\cdot) = \sum_{i=1}^N (X, \phi_i)_\mathcal{H} \phi_i(\cdot)$, for $N = \text{rank}(S)$ (possibly infinity) and $\{\phi_i\}_{i=1}^N$ the eigenvectors of S whose corresponding eigenvalues $\{\lambda_n\}_{n=1}^N$ satisfy

$$\text{E}[(X, \phi_i)_\mathcal{H}(X, \phi_j)_\mathcal{H}] = (X, S\phi_i)_\mathcal{H} = \lambda_i \delta_{ij},$$

with δ_{ij} denoting the Kronecker delta function. Alternatively, for $\lambda_i > 0$, we can write $X(\cdot) = \sum_{i=1}^N \sqrt{\lambda_i} \tilde{Z}_i \phi_i(\cdot)$ with

$$\tilde{Z}_i \equiv (X, \phi_i)_\mathcal{H} / \sqrt{\lambda_i} \tag{2}$$

uncorrelated random variables with unit variances.

The covariance kernel in (1) for $X(\cdot)$ generates a reproducing kernel Hilbert space $\mathcal{H}(K)$ (see Berlinet and Thomas-Agnan 2004) that is congruent to the Hilbert space L^2_X of all linear functionals of the X process. Specifically, L^2_X contains all finite dimensional linear combinations of $\{X(t), t \in E\}$ and their limits under the inner product $(U, V)_{L^2_X} = \text{E}[UV]$, for $U, V \in L^2_X$, while $\mathcal{H}(K) = \{f : f = \sum_{i=1}^\infty \lambda_i f_i \phi_i, \sum_{i=1}^\infty \lambda_i f_i^2 < \infty\}$. The congruence mapping $\Psi : \mathcal{H}(K) \mapsto L^2_X$ is defined by requiring every finite dimensional linear combination $\sum_{i=1}^n a_i K(\cdot, t_i)$ to map to $\sum_{i=1}^n a_i X(t_i)$, for all $a_i \in \mathbb{R}, t_i \in E$, and $n \in \mathbb{N}$. An application of the integral representation theorem of Parzen (1961) then produces the following result.

Theorem 1.1 Let $f(\cdot) = \sum_{i=1}^N \lambda_i f_i \phi_i(\cdot)$ be in $\mathcal{H}(K)$. Then,

$$\Psi(f) = \sum_{i=1}^N f_i(X, \phi_i)_{\mathcal{H}} \quad \text{and} \quad \Psi^{-1} \left(\sum_{i=1}^N f_i(X, \phi_i)_{\mathcal{H}} \right) = \sum_{i=1}^N \lambda_i f_i \phi_i, \quad (3)$$

with $\Psi^{-1} = \Psi^*$, where Ψ^* denotes the adjoint of Ψ .

For the developments here we need a further congruence that connects the closure of the image of the square root of S , denoted $S^{1/2}$, with $\mathcal{H}(K)$.

Theorem 1.2 (Eubank and Hsing 2007) The Hilbert spaces $\overline{\text{Im}(S^{1/2})} = \ker(S)^\perp$ and $\mathcal{H}(K)$ are congruent under the mapping $\Gamma : \mathcal{H} \mapsto \mathcal{H}(K)$ defined by

$$(\Gamma g)(\cdot) \equiv \sum_{i=1}^N \sqrt{\lambda_i} g_i \phi_i(\cdot), \quad (4)$$

where $g = \sum_{i=1}^N (g, \phi_i) \phi_i = \sum_{i=1}^N g_i \phi_i \in \ker(S)^\perp$. The inverse mapping

$$(\Gamma^{-1} f)(\cdot) \equiv \sum_{i=1}^N \sqrt{\lambda_i} f_i \phi_i(\cdot), \quad (5)$$

for $f = \sum_{i=1}^N \lambda_i f_i \phi_i(\cdot) \in \mathcal{H}(K)$ is also the adjoint of Γ .

Theorems 1.1 and 1.2 provide congruences that connect L_X^2 to $\mathcal{H}(K)$ and $\mathcal{H}(K)$ to $\ker(S)^\perp$. We can connect these mappings to obtain a congruence between L_X^2 and $\ker(S)^\perp \subseteq L^2(E)$. In the next section we use this to produce the natural extension of (1) to the case where X is $L^2(E)$ valued.

2 Main Result

First note that since (3)-(5) are congruences, the composition $\Omega \equiv \Psi \circ \Gamma$ is also a congruence mapping from $\ker(S)^\perp$ onto L_X^2 . Thus, if we take $f = \sum_{i=1}^N f_i \phi_i \in \ker(S)^\perp$ then,

$$\begin{aligned} \Omega(f) &\equiv \Psi \left(\Gamma \left(\sum_{i=1}^N f_i \phi_i \right) \right) = \Psi \left(\sum_{i=1}^N \frac{\lambda_i f_i \phi_i}{\sqrt{\lambda_i}} \right) \\ &= \sum_{i=1}^N \frac{f_i}{\sqrt{\lambda_i}} (X, \phi_i)_{L^2(E)} = \sum_{i=1}^N f_i \tilde{Z}_i \end{aligned} \quad (6)$$

with \tilde{Z}_i defined in (2). However, if $f \in \text{Im}(S^{1/2}) \subseteq \overline{\text{Im}(S^{1/2})} = \ker(S)^\perp$ then $f = S^{1/2}g$ for some $g \in L^2(E)$ so we may further refine (6) to write

$$\Omega(f) = \begin{cases} (X, S^{-1/2}f)_{\mathcal{H}} & \text{for } f \in \text{Im}(S^{1/2}) \text{ and,} \\ \sum_{i=1}^{\infty} f_i \tilde{Z}_i, & \text{whenever } f \in \overline{\text{Im}(S^{1/2})} \setminus \text{Im}(S^{1/2}) \text{ and } N = \infty \end{cases} \quad (7)$$

with $S^{-1/2}$ denoting the Moore-Penrose inverse of $S^{1/2}$. The inverse mapping $\Omega^{-1} : L_X^2 \mapsto \ker(S)^\perp$ for any $U = \sum_{i=1}^N f_i(X, \phi_i)_{\mathcal{H}}$ has the form

$$\begin{aligned} \Omega^{-1}(U) &= \Gamma^{-1} \left(\Psi^{-1} \left(\sum_{i=1}^N f_i(X, \phi_i)_{\mathcal{H}} \right) \right) = \Gamma^{-1} \left(\sum_{i=1}^N \lambda_i f_i \phi_i \right) \\ &= \sum_{i=1}^N \sqrt{\lambda_i} f_i \phi_i = S^{1/2}f, \end{aligned} \quad (8)$$

with $f = \sum_{i=1}^N f_i \phi_i \in \ker(S)^\perp$. We can now obtain the desired extension of (1).

Theorem 2.1 *For any $\tilde{f} = \sum_{i=1}^N \lambda_i f_i \phi_i \in \mathcal{H}(K)$ and $f = \Gamma^{-1}(\tilde{f}) \in \ker(S)^\perp$*

$$\Psi(\tilde{f}) = \begin{cases} (X, S^\dagger \tilde{f})_{\mathcal{H}} & \text{whenever } f \in \text{Im}(S^{1/2}) \text{ and,} \\ \sum_{i=1}^{\infty} (\sqrt{\lambda_i} f_i) \tilde{Z}_i & \text{for } f \in \overline{\text{Im}(S^{1/2})} \setminus \text{Im}(S^{1/2}) \text{ and } N = \infty, \end{cases} \quad (9)$$

where S^\dagger denotes the Moore-Penrose inverse of S . The inverse mapping for any $U = (X, f)_{\mathcal{H}} = \sum_{i=1}^N f_i(X, \phi_i)_{\mathcal{H}}$ with $f \in \ker(S)^\perp$ is $\Psi^{-1}(U) = Sf$.

Proof. Observe that $\Psi = \Omega\Gamma^{-1}$. Thus, for any $\tilde{f} = \sum_{i=1}^N \lambda_i f_i \phi_i \in \mathcal{H}(K)$ and $f = \Gamma^{-1}(\tilde{f}) = \sum_{i=1}^N (\sqrt{\lambda_i} f_i) \phi_i \in \ker(S)^\perp$ it follows that

$$\Psi(\tilde{f}) = \Omega(\Gamma^{-1}(\tilde{f})) = \Omega \left(\sum_{i=1}^N (\sqrt{\lambda_i} f_i) \phi_i \right) = \sum_{i=1}^N (\sqrt{\lambda_i} f_i) \tilde{Z}_i.$$

In the special case that $f \in \text{Im}(S^{1/2})$,

$$\Psi(\tilde{f}) = \Omega(\Gamma^{-1}(\tilde{f})) = (X, S^{-1/2}\Gamma^{-1}\tilde{f})_{L^2(E)} = (X, S^\dagger \tilde{f})_{L^2(E)}.$$

Similarly, $\Psi^{-1} = \Gamma\Omega^{-1}$. So, for any $U = (X, f)_{\mathcal{H}} \in L_X^2$, with $f = \sum_{i=1}^N f_i \phi_i \in \ker(S)^\perp$

$$\Psi^{-1}(U) = \Gamma \left(\Omega^{-1} \left(\sum_{i=1}^N f_i(X, \phi_i)_{\mathcal{H}} \right) \right) = \Gamma \left(\sum_{i=1}^N f_i(S^{1/2}\phi_i) \right) = Sf.$$

□

Acknowledgments. Helpful comments and suggestions by Randy Eubank are gratefully acknowledged.

References

- [1] Berlinet, A. and Thomas-Agnan, C., 2004. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, Boston.
- [2] Eubank, R. L. and Hsing T., 2007. Canonical Correlation for Stochastic Processes. *Stochastic Processes and their Application*. **118**, 16634–1661.
- [3] Laha, R. and Rohatgi V., 1979. *Probability Theory*. Wiley, New York.
- [4] Parzen, E., 1961. An Approach to Time Series Analysis. *Annals of Mathematical Statistics*. **32**, 951–989.

MULTIVARIATE CONTROL CHART FOR MONITORING THE COVARIANCE MATRIX

Edgard M. Maboudou-Tchao & Douglas M. Hawkins

University of Central Florida, Orlando, USA

&

University of Minnesota, Minneapolis, USA

Abstract : In this paper we analyze a multivariate control chart to detect small and persistent changes in the covariance matrix. We present some theoretical properties of the proposed statistic and how to use that statistic to monitor the stability of the covariance matrix of a process.

Key Words : Run length, Average run length (ARL), In-control (IC), Out-of-control (OOC)

1 Introduction

Shewhart's famous dictum suggests that proper quality control involved monitoring both the mean and the variance of the process. In the realm of multivariate process, there are several methodologies for monitoring the mean vector, such as the multivariate cumulative sum charts. Another attractive scheme is the multivariate exponentially weighted moving average (MEWMA) of Lowry et al. (1992). For monitoring the covariance matrix, the menu of choice is limited and it is this gap that we intend to fill with this paper.

2 Definition and properties of the MEWMC sequences

2.1 Definition

Suppose that the process readings are p component vectors \mathbf{X}_n , $n = 1, 2, \dots$ and that while the process is in control these process readings follow independent multivariate normal distributions with mean vector $\boldsymbol{\mu}_0$ and covariance matrix $\boldsymbol{\Sigma}_0$ which we assume are known exactly.

It is convenient to work with multistandardized data vectors rather than the 'raw' process readings \mathbf{X}_i ; for this we find a matrix \mathbf{A} with the property $\mathbf{A}\boldsymbol{\Sigma}_0\mathbf{A}' = \mathbf{I}_p$, and transform to $\mathbf{U}_i = \mathbf{A}(\mathbf{X}_i - \boldsymbol{\mu}_0)$. While in control, the \mathbf{U}_i are $N(0, \mathbf{I}_p)$. We then define the sequence \mathbf{S}_n by the recursion

$\mathbf{S}_0 = \mathbf{I}_p$, and for $n = 1, 2, \dots$

$$\mathbf{S}_n = (1 - \lambda)\mathbf{S}_{n-1} + \lambda \mathbf{U}_n \mathbf{U}_n' \quad (2.1)$$

λ is a smoothing constant such that $0 < \lambda < 1$

While the process is in-control, by decomposing \mathbf{S}_{n-1} in terms of \mathbf{S}_{n-2} , \mathbf{S}_{n-2} in terms of \mathbf{S}_{n-3} , and so on, it follows

Lemma 2.1.

$$\mathbf{S}_n = (1 - \lambda)^n \mathbf{I}_p + \lambda \sum_{k=0}^{n-1} (1 - \lambda)^k \mathbf{U}_{n-k} \mathbf{U}_{n-k}' \quad (2.2)$$

The above formula shows that \mathbf{S}_n is a linear combination of the identity matrix weighted by a coefficient $(1 - \lambda)^n$ and the random matrices $\mathbf{U}_1 \mathbf{U}_1', \dots, \mathbf{U}_n \mathbf{U}_n'$ weighted by the coefficients $\lambda(1 - \lambda)^{n-1}, \dots, \lambda$. For this reason, the sequence $\mathbf{S}_1, \dots, \mathbf{S}_n, \dots$ is called a Multivariate Exponentially Weighted Moving Covariance Matrix (MEWMC) sequence. When $\lambda \rightarrow 0$, the sequence $\mathbf{S}_1, \dots, \mathbf{S}_n$ tends to be a smoother version of the initial sequence $\mathbf{U}_1 \mathbf{U}_1', \dots, \mathbf{U}_n \mathbf{U}_n'$ and when $\lambda = 0$, then $\mathbf{S}_n = \mathbf{S}_{n-1} = \dots = \mathbf{S}_0 = \mathbf{I}_p$.

Also, we have the following theorem which states that the sequence $\mathbf{S}_1, \dots, \mathbf{S}_n, \dots$ is positive definite.

Theorem 2.2. *\mathbf{S}_n is a positive definite matrix.*

2.2 Independence of the \mathbf{S}_n

The random variables $\mathbf{U}_1, \dots, \mathbf{U}_n$ are, by definition, independent so are the random matrices $\mathbf{U}_1 \mathbf{U}_1', \dots, \mathbf{U}_n \mathbf{U}_n'$ as a consequence. However, the random matrices $\mathbf{S}_1, \dots, \mathbf{S}_n$ are not independent.

Theorem 2.3. *The sequence $\mathbf{S}_1, \dots, \mathbf{S}_n, \dots$ has a Markov property.*

This theorem is saying that the sequence $\mathbf{S}_1, \dots, \mathbf{S}_n, \dots$ is a Markov chain (nonstationary)

2.3 Expectations of the MEWMC

By using lemma 2.1, we can show using an argument by induction that

Lemma 2.4.

$$E(\mathbf{S}_n) = \mathbf{I}_p \quad (2.3)$$

Similarly, the covariance matrix of \mathbf{S}_n is given in the following lemma

Lemma 2.5.

$$Cov(\mathbf{S}_n) = \frac{2\lambda}{2 - \lambda} [1 - (1 - \lambda)^{2n}] (\mathbf{I}_p \otimes \mathbf{I}_p). \quad (2.4)$$

As $n \rightarrow \infty$, $Cov(\mathbf{S}_n) = \frac{2\lambda}{2 - \lambda} (\mathbf{I}_p \otimes \mathbf{I}_p)$. So, the asymptotic covariance of the process \mathbf{S}_n is $\frac{2\lambda}{2 - \lambda} (\mathbf{I}_p \otimes \mathbf{I}_p)$

3 Monitoring the covariance matrix

We define the MEWMC statistic to compare the matrix \mathbf{S}_n with the identity. Therefore, we use the statistic

$$c_n = \text{tr}(\mathbf{S}_n) - \log|\mathbf{S}_n| - p \tag{3.1}$$

where tr is the trace of a matrix, $|\mathbf{S}_n|$ is the determinant of \mathbf{S}_n .

The control chart use then consists of plotting c_n against n , and signaling a loss of control if $c_n > h$ where the control limit h is chosen to achieve a specified in-control ARL. The exact distribution of c_n not available so the control limits h are obtained via simulation.

4 Example

In this section, we illustrate the application of the MEWMC chart for covariance shifts so that scatter shifts can be detected. The data set used was kindly provided to us by Dr Franz Halberg. In this work, subjects are equipped with instruments that measure and record physiological variables. The wearer’s blood pressure and heart rate were measured and recorded every 15 minutes for 6 years. Prior to analysis using SPC methods, each week’s raw data are condensed into weekly summary numbers which include:

- SBP: A mean systolic blood pressure
- DBP: A mean diastolic blood pressure
- HR: A mean of heart rate
- MAP: An overall mean arterial pressure

We set the smoothing constant λ to 0.1 and the in-control average run length (IC ARL) to 500. Hence the control limit is $h = 1.342$ for the MEWMC. The MEWMC chart signals an out-of-control (OOC) behavior at observation 22 (figure 1 in the appendix), so the shift apparently came around reading 5. With these clues, we go back and diagnose. We found out that the residual variances of MAP and HR have changed. The other variables seem fine.

5 Conclusion

We define a new process, MEWMC. We establish some theoretical properties and use it to set up a method to monitor the covariance matrix of a process.

Some References

1 Chambers, J. M., (1971). Regression Updating. *Journal of the American Statistical Association*, **66**, 744–748

- 2 Eaton, M. L.(1983) *Multivariate Statistics: A vector approach*. John Wiley and Sons New York.
- 3 Hawkins, D. M., Maboudou-Tchao, E. M., 2008 Multivariate exponentially weighted moving covariance matrix. *Technometrics*, 50, 155-166
- 4 Hawkins, D. M. and Eplett, W. J. R. (1982) “The Cholesky Factorization of the inverse correlation or covariance matrix in multiple regression”. *Technometrics*. 24, 191-198.
- 5 Montgomery, D. C. (2005), *Introduction to Statistical Quality Control*. 5th edition, Wiley

Appendix

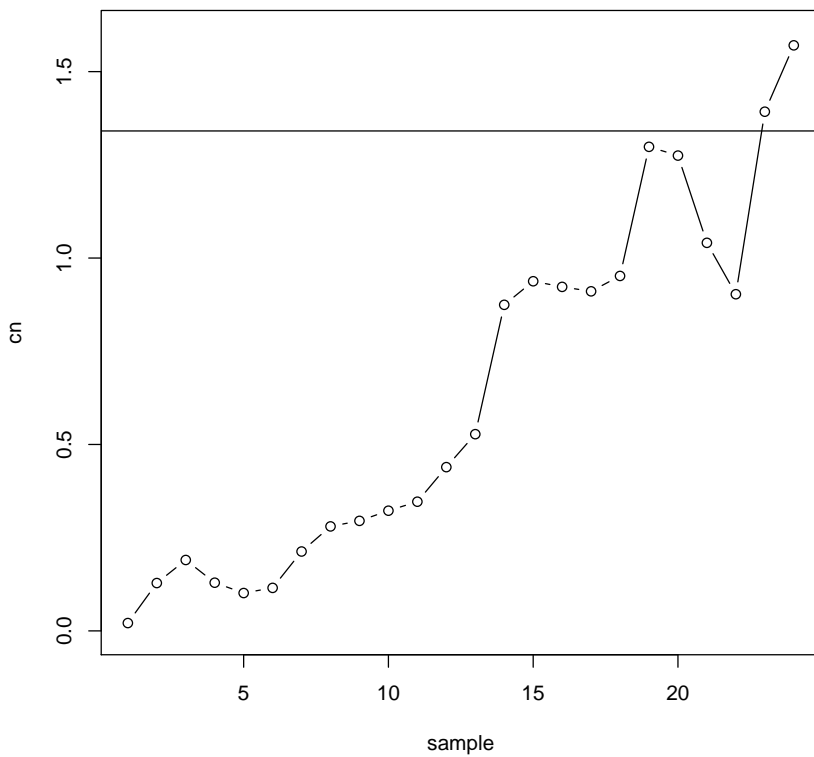


Figure 1: MEWMC for the Ambulatory data

A Note on the Probability of Correct Selection for Large k Populations, with Application to Microarray Data

Jason Wilson
jason.wilson@biola.edu
Department of Mathematics
Biola University
13800 Biola Ave.
La Mirada, CA 90639

August 17, 2009

1 Introduction

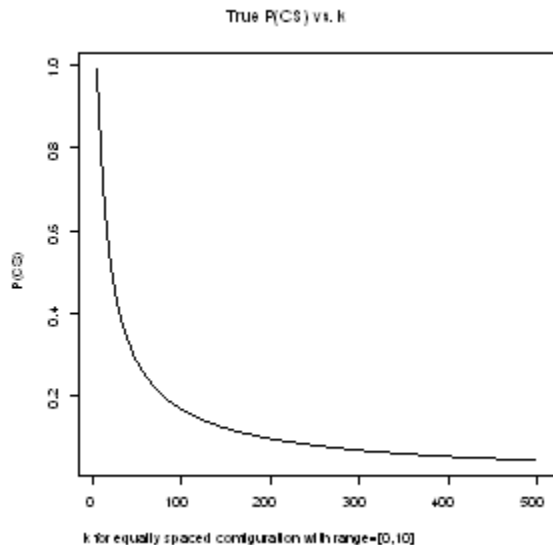
Ranking and Selection Methodology (RSM), also sometimes referred to as Ranking and Selection Procedures, is a well developed branch of Statistics, complementary to but distinct from the more widely known branch called Hypothesis Testing (1200+ papers by 1982, see Dudewicz and Koo 1982). In the last century, data sets evolved from $k = 1$ population, to including small k ($\approx 10^1$), large k ($\approx 10^2$), and even recently extremely large k ($\approx 10^3$ or 4) populations. Multiple hypothesis testing procedures have been continually developed to keep up with growing k . Formal RSM, however, still only tends to be employed on small k problems. In this research, I hope to encourage and facilitate future work on RSM for larger k by reviewing the literature on one lesser known aspect of RSM called 'probability of correct selection' (PCS). I believe PCS holds promise for obtaining additional useful information from today's extremely large k data sets.

PCS is the probability of making a correct selection of the *best* t out of k populations (CS_t), according to a given decision rule. It is therefore measuring the likelihood that the specified selection goal is achieved using the given decision rule. For example, if PCS is 0.93, then there is a 93% chance that the selection goal is achieved using the given decision rule. Estimating the PCS of the best one ($t = 1$) population was introduced by Olkin, Sobel and Tong (1982) within the Indifference Zone framework of RSM. In my review, I explored a body of literature on the point estimators of $P(CS_{t=1})$ for small k populations (Cui and Wilson 2008a). In my main paper (Cui and Wilson 2008b), $P(CS_{t=1})$ was extended for $t > 1$, and two tuning parameters, d-best and G-best correct selection were proposed, with examples. In what follows, I provide a brief sketch of the motivation and application of PCS for large k problems.

2 Motivation

In practice, the range of the parameters of probability distributions tends to be bounded. Consider k location parameter populations and suppose we want to select the population with the largest location parameter. Therefore, the range of all k location parameters will tend to be bounded in practice. Suppose we sample from the k populations and estimate the k population parameters from the samples. When k is small, it is easy to think that the population with the largest sample statistic will be set apart and correctly indicate the population with the largest population parameter. However, when k becomes increasingly large, it becomes less likely that the largest location parameter population will stand out. The figure below illustrates what happens when k parameters are uniformly spaced on $[0, 10]$.

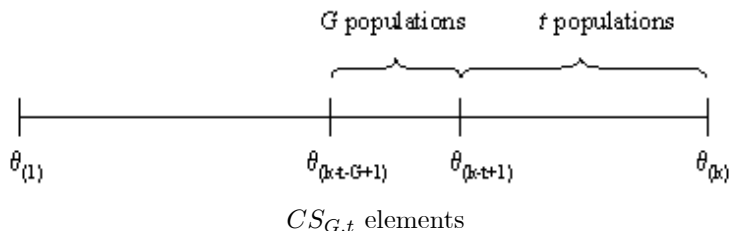
For realistic applications, when k is large, it is often of interest to select $t > 1$. Nevertheless, the problem of decreasing PCS persists. However, by introducing an appropriate tuning parameter, PCS can be increased. There are at least three reasons for introducing a tuning parameter: (i) for large k , $P(CS_t)$



may be too small to be useful; (ii) for large k , often the goal is a screen, as opposed to significance tests; and (iii) such PCS may be used to compare competing multiple testing procedures ("meta-method"). This paper only focuses on reason (i) (but see Cui and Wilson 2008b)

3 Application

Define $CS_{G,t}$ as when the top $G + t$ statistics select a set including the top t parameters. We refer to selection $CS_{G,t}$ as "G-best".



In the diagram, $\theta_{(k)}$ refers to the k th population order statistic. The selection rule is to select the populations according to the top $G + t$ sample order statistics $Y_{[k-t-G+1]}, \dots, Y_{[k]}$. See Mahamunulu (1968).

The formula and method for calculating $P(CS_{G,t})$ are given in Cui and Wilson (2008b). Some of the features of the method are: (i) applications can yield useful probabilities; (ii) estimator (under normal assumption) is consistent and asymptotically normal (proved, but not yet published); (iii) estimation is computationally intensive; (iv) no unknown errors (i.e. type II error; you either are correct or not); and (v) developed for the following test statistic distributions: normal, Student's t , and unknown (see Cui and Wilson 2009). Simulation studies reveal that the estimator is moderately robust to nonnormality, but is biased upwards for small t ($t \leq 10$).

To illustrate, I will consider the Apo AI experiment (Callow, et. al. 2000). In it, cDNA microarrays were used to measure gene expression in the livers of 8 inbred control mice versus 8 mice with the Apo AI gene "knocked out". The goal was to identify genes with altered expression in the livers of these two groups of mice (Dudoit, et. al. 2003). Welch two-sample t -statistics were used on the $k = 6383$ genes with max T permutation test adjusted p -values shown.

top t	rawp	t-stat	st error	maxT	$\hat{P}(CS_t)$	$\hat{P}(CS_{2,t})$
1	7.33e-07	16.50	0.197	0.0002	1.000	1.00
2	2.46e-05	9.79	0.102	0.0008	0.529	0.94
3	3.55e-05	9.26	0.193	0.0009	0.431	0.89
4	4.99e-05	8.78	0.338	0.0012	0.467	0.98
5	1.00e-04	7.89	0.124	0.0056	0.379	0.94
6	1.03e-04	7.84	0.125	0.0061	0.632	0.80
7	5.94e-04	5.91	0.156	0.0664	0.173	0.37
8	7.41e-04	5.69	0.129	0.0889	0.051	0.12
9	1.31e-03	5.16	0.193	0.1771	0.008	0.03

At $\alpha = 0.10$, $t = 8$ genes may be selected. This is the number selected in Dudoit, et al. 2003). However, the probability that those 8 genes are the true top 8 out of the 6383 is approximately 0.051. If one keeps $t = 8$, but selects an extra two genes, the probability increases to approximately 0.12 that the 10 genes selected contains the top 8. This is not much improvement. However, if one holds the total number selected at 8, $t + G = 8$, and only requires that 6 of those 8 be in the true top 8, then the probability jumps to 0.80. The reason for this dramatic improvement is that there is a big gap between the t-statistics of the 6th (7.84) and 7th (5.91) top t-statistics. This does not say that genes 7 and 8 are somehow unlikely to be statistically significant; indeed, they are at the 10% level. Rather, it simply says that the probability that the sample top 8 are unlikely to be the true top 8, even though they may be statistically significant. In other words, in limiting the conclusion to 8, one is probably including some non-top-8 genes. $P(CS_{2,6})$ is a step towards quantifying this uncertainty.

4 Conclusion

PCS is offered as an extension of $P(CS_t)$ from Ranking and Selection literature, for large k applications. Possible areas include: neuroimaging, financial data, and fingerprints. Further work includes addressing population dependence.

Acknowledgement

Thank you to Xinping Cui, my dissertation advisor, for collaborating with me in this research.

LABELLING SWITCHING FOR BAYESIAN MIXTURE MODELS

Weixin Yao & Bruce Lindsay

Kansas State University, Manhattan, KS, USA

ℰ

The Pennsylvania State University, University Park, PA, USA

Abstract : A fundamental problem for Bayesian mixture model analysis is label switching, which occurs due to the non-identifiability of the mixture components under symmetric priors. In this article, we propose two labelling methods to solve this problem. The first method, denoted by PM(ALG), is to explore the geometry of the mixture posterior by using each MCMC draw as a starting point for the ascent algorithm ALG and labelling the samples based on the modes of the posterior density they converge to. The natural assumption here is that *the samples converged to the same mode should have the same labels*. The PM(ALG) labelling method has some computational advantages over other popular labelling methods. The second method does labelling by maximizing the normal likelihood of the labelled Gibbs samples based on the asymptotic normality for the posterior.

Key Words : Label switching; Bayesian approach; Markov chain Monte Carlo; Mixture model; Posterior modes

1 Introduction

Suppose $\mathbf{x} = (x_1, \dots, x_n)$ are independent observations from a m -component mixture density

$$p(x; \boldsymbol{\theta}) = \pi_1 f(x; \lambda_1) + \pi_2 f(x; \lambda_2) + \dots + \pi_m f(x; \lambda_m),$$

where $\boldsymbol{\theta} = (\pi_1, \dots, \pi_m, \lambda_1, \dots, \lambda_m)^T$, $f(\cdot)$ is some parametric component density/mass function, λ_j is the component specific parameter, which can be scalar or vector, and π_j is the proportion of j^{th} component with $\sum_{j=1}^m \pi_j = 1$. The likelihood for \mathbf{x} is

$$L(\boldsymbol{\theta}; \mathbf{x}) = \prod_{i=1}^n \{\pi_1 f(x_i; \lambda_1) + \pi_2 f(x_i; \lambda_2) + \dots + \pi_m f(x_i; \lambda_m)\}. \quad (1)$$

For any permutation $\omega = (\omega(1), \dots, \omega(m))$ of the identity permutation $(1, \dots, m)$, define the corresponding permutation of the parameter vector θ by

$$\theta^\omega = (\pi_{\omega(1)}, \dots, \pi_{\omega(m)}, \lambda_{\omega(1)}, \dots, \lambda_{\omega(m)})^T.$$

Noticing that $L(\theta^\omega; \mathbf{x})$ is numerically the same as $L(\theta; \mathbf{x})$ for any permutation ω , hence if $\hat{\theta}$ is the maximum likelihood estimator (MLE), $\hat{\theta}^\omega$ is the MLE for any permutation ω . This is so-called label switching problem.

The label switching problem also occurs in Bayesian mixtures. If we do not have prior information that distinguishes between the components of a mixture model, then the posterior is invariant under the relabelling of mixture components and thus the posterior has $m!$ symmetric modal regions. Note that the marginal posterior distributions for the parameters will be also identical for each mixture component. Hence, it is meaningless to draw inference, relating to individual components, directly from Markov chain Monte Carlo (MCMC) samples using ergodic averaging before solving the label switching problem. Given the MCMC samples $(\theta_1, \dots, \theta_N)$, our aim is to find the labels $(\omega_1, \dots, \omega_N)$ such that $\theta_1^{\omega_1}, \dots, \theta_N^{\omega_N}$ are all in the same modal region, i.e. have the same label meaning.

The easiest way to solve the label switching is to use an explicit parameter constraint so that only one permutation can satisfy it. See Diebolt and Robert (1994) and Richardson and Green (1997), for example. Another popular labelling method is relabelling algorithm (Celeux 1998; Stephens 2000), which is based on minimizing a Monte Carlo risk. In the following section, we will briefly introduce two new labelling methods. For more detail, please refer to Yao and Lindsay (2009).

2 Methodology

2.1 Labelling Using Modal Clusters

If $\tilde{\theta}$ is a mode, then so is $\tilde{\theta}^\sigma$ for any permutation σ . If an ascending algorithm ascends from θ to $\tilde{\theta}^\sigma$, we will say θ has the same labelling as $\tilde{\theta}^\sigma$. If the algorithm is permutation symmetric we will also know that $\theta^{\sigma^{-1}}$, where σ^{-1} is the inverse permutation of σ such that $(\theta^\sigma)^{\sigma^{-1}} = \theta$ for any θ and σ , will be given the same labelling as $\tilde{\theta}$.

If the posterior density has a maximal mode at $\tilde{\theta}$, it also has maximal modes at all permutations of $\tilde{\theta}$. We can pick one such mode to be our reference mode, say by order constraint labelling on some parameter. Denote by $\hat{\theta}$ the chosen reference maximal mode. When a sampled θ is used as a starting point for the chosen ascending algorithm, if it converges to a maximal mode, say $\hat{\theta}^\sigma$, then the natural label of θ is σ^{-1} since $\theta^{\sigma^{-1}}$ would ascend to $\hat{\theta}$. If the θ converges to a minor mode, say θ_* , we could create a labelling system for all the samples θ that are attracted to θ_* (or its permutations) by creating a secondary reference mode $\hat{\theta}_2$. If the reference mode $\hat{\theta}_2$ was chosen so that it matched the label with

the maximal mode $\hat{\theta}$ using a risk based criterion that makes $\hat{\theta}_2 = \theta_*^\sigma$ most similar to $\hat{\theta}$ for some σ , then we have a system that labels all points attracted to both the maximal and minor modes. One can extend this idea to any number of minor modes. The algorithm of our proposed labelling method, given the reference mode $\hat{\theta}$, is as follows .

Algorithm 1: *Labelling based on posterior modes and an ascent algorithm (PM(ALG))*

Step 1: Taking each MCMC sample $\{\theta_t, t = 1, \dots, N\}$ as the initial value, find the corresponding converged mode $\{\mathbf{m}_t, t = 1, \dots, N\}$ using the given ascent algorithm ALG.

Step 2: Apply to m_t the order constraint labelling used to define $\hat{\theta}$, denoted by σ_t^* (hence $m_t^{\sigma_t^*}$ has the same order constraint as $\hat{\theta}$) and find the label σ_t of θ_t based on the following situations.

a) If $\mathbf{m}_t^{\sigma_t^*}$ is $\hat{\theta}$, up to numerical error, then $\sigma_t = \sigma_t^*$.

b) If $\mathbf{m}_t^{\sigma_t^*}$ is not $\hat{\theta}$, but it is equivalent (up to a permutation) to a known reference minor mode, say $\hat{\theta}_2$, assign the label σ_t such that $\mathbf{m}_t^{\sigma_t} = \hat{\theta}_2$.

c) If $\mathbf{m}_t^{\sigma_t^*}$ is not $\hat{\theta}$ and is not equivalent to a preexisting reference minor mode, create a new reference minor mode $\mathbf{m}_t^{\sigma_t}$, where σ_t is based on a risk based criterion such as least squares:

$$\sigma_t = \arg \min_{\sigma} (\mathbf{m}_t^{\sigma} - \hat{\theta})^T (\mathbf{m}_t^{\sigma} - \hat{\theta}). \quad \square \quad (2)$$

Yao and Lindsay (2009) introduced an ECM-type ascent algorithm to find the reference mode $\hat{\theta}$ and to locate other posterior modes.

There are several nice properties of the PM(ALG) method. Firstly, unlike a typical relabelling algorithm, the PM(ALG) method gives an answer that does not depend on a set of initial labels, the choice of which can change the labelling. Secondly, the PM(ALG) method is an online algorithm and it can do labelling along with the MCMC sampling process. Hence the storage requirements are reduced. Finally, the PM(ALG) method does not require one to compare $m!$ permutations when doing labelling except for the minor modes. This property can make PM(ALG) much faster than some other labelling methods when m is large.

2.2 The Classification MLE Method

From the asymptotic theory for the posterior distribution, see Walker (1969) and Frühwirth-Schnatter (2006, sec. 1.3, 2.4.3, 3.3), we know that when sample size is large, the ‘‘correctly’’ labelled MCMC samples should, approximately, follow the normal distribution. Based on this property, we propose another method to do labelling based on minimizing the following negative log normal likelihood over $(\bar{\theta}, \Sigma, \sigma)$,

$$L(\bar{\theta}, \Sigma, \sigma) = N \log(|\Sigma|) + \sum_{t=1}^N (\theta_t^{\sigma_t} - \bar{\theta})^T \Sigma^{-1} (\theta_t^{\sigma_t} - \bar{\theta}) \quad (3)$$

where $\bar{\theta}$ is the center value for the normal distribution, Σ is the covariance structure, and $\sigma = (\sigma_1, \dots, \sigma_N)$.

The algorithm to find labels by minimizing (3) is as follows.

Algorithm 3: *Labelling by normal likelihood (NORMLH)*

Starting with some initial values for $(\sigma_1, \dots, \sigma_N)$ (setting them based on an order constraint, for example), iterate the following two steps until a fixed point is reached.

Step 1: Update $\bar{\theta}$ and Σ by minimizing (3)

$$\bar{\theta} = \frac{1}{N} \sum_{t=1}^N \theta_t^{\sigma_t},$$

$$\Sigma = \frac{1}{N} \sum_{t=1}^N (\theta_t^{\sigma_t} - \bar{\theta})(\theta_t^{\sigma_t} - \bar{\theta})^T.$$

Step 2: For $t = 1, \dots, N$, choose σ_t by

$$\sigma_t = \arg \min_{\sigma} (\theta_t^{\sigma} - \bar{\theta})^T \Sigma^{-1} (\theta_t^{\sigma} - \bar{\theta}). \quad \square$$

The NORMLH method has a simple and nice explanation and runs much faster than the PM(ALG) method if the number of components m is not large. As Yao and Lindsay (2009) pointed out, the NORMLH performed somewhat better than the other labelling methods at recreating the PM(ALG) labels.

References

- Celeux, G.(1998), “Bayesian inference for mixtures: The label switching problem,” In *Compstat 98-Proc. in Computational Statistics* (eds. R. Payne and P.J. Green), 227-232. Physica, Heidelberg.
- Diebolt, J. and Robert, C. P. (1994), “Estimation of finite mixture distributions through Bayesian sampling,” *Journal of the Royal Statistical Society*, B56, 363-375.
- Frühwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models*, Springer.
- Richardson, S. and Green, P. J. (1997), “On Bayesian analysis of mixtures with an unknown number of components” (with discussion), *Journal of Royal Statistical Society*, B59, 731-792.
- Stephens, M.(2000), “Dealing with label switching in mixture models,” *Journal of the Royal Statistical Society*, B62, 795-809.
- Walker, A. M. (1969), “On the asymptotic behaviour of posterior distributions,” *Journal of the Royal Statistical Society*, B31, 80-88.
- Yao, W. and Lindsay, B. G. (2009), ”Bayesian mixture labelling by posterior density,” *Journal of American Statistical Association*, 104, 758-767.

**ON THE EXISTENCE OF THE MAXIMUM LIKELIHOOD
ESTIMATOR FOR GAUSSIAN GRAPHICAL MODELS AND A
NEW GRAPH INVARIANT**

AHMAD S. YASAMIN
SAMSI

ABSTRACT. In this paper we discuss the smallest size of a sample data, taken from a Gaussian graphical model, which guarantees the existence of the maximum likelihood estimator.

1. INTRODUCTION.

In classical multivariate statistics it is well-known that if the size of a sample data taken from a multivariate Gaussian model is larger than the dimension of the underlying random vector, then the existence of the **MLE**, with probability one, is guaranteed. For a general Gaussian model sufficient and necessary conditions are given in terms of the so-called *treewidth* and *clique* number of the graph representing the model. In this expository paper we express this quantity as a graph parameter and then, methodically, investigate its properties.

2. PRELIMINARY NOTIONS

Let $A = (a_{ij})$ be a matrix indexed by a finite set V . If $\alpha, \beta \subset V$, then $A[\alpha, \beta]$ will denote the submatrix $(a_{ij} : i \in \alpha, j \in \beta)$. If $\alpha = \beta$, then the submatrix is called a principal submatrix of A and is denoted by $A[\alpha]$, or A_α . We denote $\mathbf{S}_n(\mathbb{R})$, $\mathbf{PSD}_n(\mathbb{R})$, and $\mathbf{PD}_n(\mathbb{R})$ for the set of n -by- n symmetric matrices, semi-definite matrices and positive definite matrices, respectively.

Let $\mathcal{G}=(V, E)$ be a simple undirected graph with n vertices, unless otherwise stated. The the set of cliques, i.e. complete subgraphs of \mathcal{G} , is denoted by $\mathcal{C}(\mathcal{G})$. For a matrix $A \in \mathbf{M}_n(\mathbb{R})$ the *strong rank* of A , denoted by $\bar{r}(A)$, is the largest positive integer r such that each r -by- r principal sub-matrix of A is non-singular. We write $A \stackrel{\mathcal{G}}{=} B$ for two matrices $A = (a_{ij}), B = (b_{ij}) \in \mathbf{M}_n(\mathbb{R})$ if

$$a_{ij} = b_{ij} \text{ whenever } (i, j) \in E, \text{ or } i = j.$$

- Let $A \in \mathbf{PSD}_n(\mathbb{R})$. The set $\mathcal{V}(\mathcal{G}, A)$ is defined to be

$$\{M \in \mathbf{PSD}_n(\mathbb{R}) : M \stackrel{\mathcal{G}}{=} A\}.$$

- The function $\rho_{\mathcal{G}} : \mathbf{PSD}_n(\mathbb{R}) \rightarrow \mathbb{Z}^+$ is defined by

$$\rho_{\mathcal{G}}(A) = \max\{\text{rank}(M) : M \in \mathcal{V}(\mathcal{G}, A)\}.$$

- Let $H(\mathcal{G}, k)$ be the set of all $A \in \mathbf{PSD}_n(\mathbb{R})$ with $\bar{r}(A) \geq k$.

Date: August 16, 2009

2000 Mathematics Subject Classification. 62-05, 15.

I'd like to thank my mentor Seth Sullivant in SAMSI for introducing this problem to me.

Definition 2.1. The *Gaussian rank* of \mathcal{G} , $r(\mathcal{G})$, to be the smallest positive integer r such that

$$\rho_{\mathcal{G}}(A) = n \quad \text{for each } A \in H(\mathcal{G}, r).$$

In other words, the Gaussian rank of \mathcal{G} is the smallest integer r with this property:

For each n -by- n positive semi-definite matrix A , if the strong rank of A is r , then there exists a positive definite matrix $M \in \mathbf{PD}_n(\mathbb{R})$ such that $A \stackrel{\mathcal{G}}{=} M$.

Remark 2.1. The Gaussian rank of a graph \mathcal{G} is indeed equal to the size of the smallest data set, taken from a Gaussian graphical model represented by \mathcal{G} , that guarantees the existence of the **MLE**.

Example 2.1.

- (1) For the complete graph K_n we have $r(K_n) = n$.
- (2) More generally, If \mathcal{G} is chordal, i.e. a graph with no induced cycle of length greater than or equal to 4, then $r(\mathcal{G}) = \omega(\mathcal{G})$, where $\omega(\mathcal{G})$ is the size of largest complete subgraph of \mathcal{G} .
- (3) For a p -cycle C_p , i.e a cycle of length p , we have $r(C_p) = 3$ (see [1]).
- (4) More generally, if \mathcal{G} is a series-parallel graph, i.e. a graph with no K_4 minor, then $r(\mathcal{G}) = 3$.

3. SOME PROPERTIES OF GAUSSIAN RANK

In this section we list some of the properties of Gaussian rank. The main reason for studying these properties are to find a practical method, or algorithm, for computing the Gaussian rank of a given graph.

- (1) If \mathcal{G}_1 is a subgraph of \mathcal{G} , then $r(\mathcal{G}_1) \leq r(\mathcal{G})$.

Proof. Let $|\mathcal{G}_1| = m$, $r = r(\mathcal{G})$, and suppose $m > r$, otherwise there is nothing to prove. if $A \in H(m, r)$, then for

$$\tilde{A} = \begin{pmatrix} A & \mathbf{0} \\ \mathbf{0} & I_{n-m} \end{pmatrix}$$

we have $\bar{r}(\tilde{A}) = r$ and, therefore there exists a matrix $M \in \mathbf{PD}_n(\mathbb{R})$ such that $M \stackrel{\mathcal{G}}{=} \tilde{A}$. Thus $M_{[m]}$, the principal sub-matrix of M corresponding to indices $[m] \times [m]$, is in $\mathbf{PD}_m(\mathbb{R})$ and $M_{[m]} \stackrel{\mathcal{G}_1}{=} A$. \square

- (2) Let \mathcal{G} be the clique sum of two subgraphs $\mathcal{G}_1 = (V_1, E_1)$ and $\mathcal{G}_2 = (V_2, E_2)$, i.e. $\mathcal{G} = \mathcal{G}_1 \cap \mathcal{G}_2$ and the induced subgraph of $V_1 \cap V_2$ in \mathcal{G} is a clique in both \mathcal{G}_1 and \mathcal{G}_2 . We write this as

$$\mathcal{G} = \mathcal{G}_1 \oplus_k \mathcal{G}_2,$$

where k is the size of the common clique in \mathcal{G}_1 and \mathcal{G}_2 . Then we have

$$r(\mathcal{G}) = \max\{r(\mathcal{G}_1), r(\mathcal{G}_2)\}.$$

Proof. Without loss of generality, we may assume $V_1 = \{1, \dots, m\}$ and $V_2 = \{m - k, \dots, n\}$. By property (1), we already have

$$r(\mathcal{G}) \geq \max\{r(\mathcal{G}_1), r(\mathcal{G}_2)\}.$$

To show the opposite side of this inequality let $A \in H(\mathcal{G}, r)$, where $r = \max\{r(\mathcal{G}_1), r(\mathcal{G}_2)\}$. We partition the matrix A according to

$$\begin{pmatrix} B_1 & B_{12} & \star \\ B_{12}^T & C & D_{12} \\ \star & D_{12}^T & D_2 \end{pmatrix},$$

where

$$B = \begin{pmatrix} B_1 & B_{12} \\ B_{12}^T & C \end{pmatrix} \in \mathbf{PPD}_m(\mathcal{G}_1), \quad D = \begin{pmatrix} C & D_{12} \\ D_{12}^T & D_2 \end{pmatrix} \in \mathbf{PPD}_{n-m}(\mathcal{G}_2), \quad \text{and } C \in \mathbf{PD}_k(\mathbb{R}).$$

Since $\bar{r}(A) = r$ we have $\bar{r}(B)$ and $\bar{r}(D)$ are at least equal to r and, therefore there are

$$P = \begin{pmatrix} P_1 & P_{12} \\ P_{12}^T & C \end{pmatrix} \in \mathbf{PD}_m(\mathbb{R}) \quad Q = \begin{pmatrix} C & Q_{12} \\ Q_{12}^T & Q_2 \end{pmatrix} \in \mathbf{PD}_{n-m}(\mathbb{R}),$$

such that $P \stackrel{\mathcal{G}_1}{\cong} B$ and $Q \stackrel{\mathcal{G}_2}{\cong} D$. We have the partial positive definite matrix

$$\begin{pmatrix} P_1 & P_{12} & \star \\ P_{12}^T & C & Q_{12} \\ \star & Q_{12}^T & Q_2 \end{pmatrix}.$$

corresponds to a chordal graph and thus, by [2], it has a positive definite completion, say M . Clearly, $M \stackrel{\mathcal{G}}{\cong} A$. \square

- (3) Let $G = (V, E)$ with $V = \{v, 1, \dots, n\}$. Then we have

$$r(G - v) \leq r(\mathcal{G}) \leq r(\mathcal{G} - v) + 1.$$

moreover, $r(\mathcal{G}) = r(\mathcal{G} - v) + 1$ if $d_G(v) = n$, i.e v is adjacent to all other vertices.

Proof. Let set $r_v = r(\mathcal{G} - v)$. We need to show that $r(\mathcal{G}) \leq r_v + 1$. If $r_v = n$, then there is nothing to prove. Assume otherwise and let $\tilde{A} \in H(\mathcal{G}, r_v + 1)$. Note that without loss of generality we may assume that $w_i = 0$ if $(v, i) \notin E$. We partition \tilde{A} as

$$\begin{pmatrix} \lambda & \mathbf{w}^T \\ \mathbf{w} & B \end{pmatrix},$$

where $\lambda > 0$, $\mathbf{w} \in \mathbb{R}^n$ and $B \in H(\mathcal{G} - v, r_v + 1)$. Therefore

$$A = B - \frac{1}{\lambda} \mathbf{w} \mathbf{w}^T \in H(\mathcal{G} - v, r_v).$$

So by definition, there is a matrix $M \in \mathbf{PD}_n(\mathbb{R})$ such $M \stackrel{\mathcal{G}-v}{=} A$. We have

$$\tilde{M} := \begin{pmatrix} \lambda & \mathbf{w}^T \\ \mathbf{w} & M + \frac{1}{\lambda} \mathbf{w} \mathbf{w}^T \end{pmatrix} \in \mathbf{PD}_{n+1}(\mathbb{R}),$$

and $\tilde{M} \stackrel{\mathcal{G}}{=} \tilde{A}$. □

- (4) If \mathcal{G} is a *subdivision* of \mathcal{G}_1 , i.e. it is obtained from \mathcal{G}_1 by replacing some of the edges with independent paths, then $r(\mathcal{G}) \leq r(\mathcal{G}_1)$.

Proof. Let $e = (u, v)$ be an edge in E_1 . Without loss of generality, we assume that \mathcal{G} is obtained from \mathcal{G}_1 by adding an additional vertex w to V_1 and two edges $(w, v), (w, u)$ to E_1 and deleting e . Let $\mathcal{G}_2 = \mathcal{G} + e$, i.e. $V_2 = V \cup \{e\}$. Thus we have $\mathcal{G}_2 = \mathcal{G}_1 \oplus_2 K_3$, where the vertex set of K_3 is $\{u, v, w\}$. This shows that $r(\mathcal{G}_2) \leq tw(\mathcal{G}_2) = \max(r(\mathcal{G}_1), 3)$. So unless $r(\mathcal{G}_1) \leq 2$ we have $r(\mathcal{G}) \leq r(\mathcal{G}_1)$. It is easy to check directly that inequality still holds if $r(\mathcal{G}_1) \leq 2$. □

- (5) If \mathcal{G} is a subgraph of the complete bipartite graph $K_{n,m}$, then $r(\mathcal{G}) \leq \min\{m, n\} + 1$.

Proof. Let $X, Y \subset V$ with $|X| = m, |Y| = n$ and $E_{n,m} = \{(u, v) : u \in X, v \in Y\}$. Suppose $m \leq n$. Let $\tilde{K}_{n,m}$ be the graph with the same vertex set as $K_{n,m}$ and the set of edges $\tilde{E}_{n,m} = E_{n,m} \cup \{(u, w) : u, w \in X\}$. Then $\tilde{K}_{n,m}$ is triangulated and $\omega(\tilde{K}_{n,m}) = m + 1$. Therefore $r(\mathcal{G}) \leq r(K_{n,m}) \leq r(\tilde{K}_{n,m}) = m + 1$. □

4. TWO OPEN QUESTIONS

The *clique contraction number*, or *Hadwiger number*, of a graph \mathcal{G} , denoted by $\eta(\mathcal{G})$, is the size of the largest clique that one can get by contracting some of the edges of \mathcal{G} . For any graph \mathcal{G} we conjecture that

- (1) $r(\mathcal{G}) \leq \eta(\mathcal{G})$,
- (2) $r(\mathcal{G}) \geq \delta(\mathcal{G}) = \min\{d_{\mathcal{G}}(v) : v \in V\}$.

REFERENCES

- [1] Buhl S. L. (1993). *On the existence of maximum likelihood estimators for graphical Gaussian models*. Scandinavian Journal of Statistics, 1993, vol. 20, no3, pp. 263-270.
- [2] Grone R, Johnson C. R., Sa E. M. and Wolkowicz H. *Positive definite completions of partial Hermitian matrices*. Linear Algebra and Its Applications 58 (1984), 109-124.
- [3] Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press, 1996.

19 TW ALEXANDER DRIVE, DURHAM, NC 27703
E-mail address: `syasamin@samsi.info`

NONLINEAR JOINT MEAN-COVARIANCE MODELING FOR LONGITUDINAL DATA ANALYSIS

Peng Zhang & Peter X.-K. Song

University of Alberta, Edmonton, Canada

&

University of Michigan, Ann Arbor, USA

Abstract : We propose a generalization of mixed effects models by directly assuming the mean vector itself to be random in a longitudinal data analysis. The resulting model is termed the random mean model, which includes the mixed effects model as a special case. The generalization allows us to relax the restriction that random effects have to be a subset of fixed effects. It is capable of modeling the mean and covariance matrix jointly where Cholesky decomposition of the inverse of the covariance matrix makes the parameterisation unconstrained. In this paper, we focus on a nonlinear mixed effects random mean model at which we exhibit some appealing features of the proposed model in the longitudinal data of high heterogeneity. Although predicting subject-specific effects becomes impossible the random mean model allows much more flexible covariance structures which is desirable in some situations with complex longitudinal data.

Key Words : Random mean models; Mixed effects models; Nonlinear regression; Longitudinal data analysis.

1 Introduction

Longitudinal data often exhibit a relationship between response and explanatory variables that is best characterized by a model nonlinear in its parameters. The appropriate nonlinear model may be derived theoretically or empirically. For example the kinetics of drug or other substances is often represented by some functions nonlinear in parameters. In this paper we present one study, which recorded the decay of intraocular gas (C_3F_8) in complex retinal surgeries following initial injection in an ophthalmology study, reported in Meyers *et al.* (1992). The outcome variable was the percent of gas left in the eye. The gas, with three different concentration levels, 15%, 20% and 25%, was injected into the eye before surgery for 31 patients. They were then followed three to eight (average of 5) times over a three-month period, and the volume of gas in the eye at the follow-up

times were recorded as a percentage of the initial gas volume. The primary interest was to investigate whether concentration levels of the gas injected in patients' eyes affect the decay rate of the gas.

The volume of gas left in the eye decay through the follow up period due to natural absorption. The decay rate of the gas is proportional to the volume remaining in the eye. Hence, the decay rate obeys the differential equation:

$$\frac{df}{dt} = -\kappa(\alpha - f) \quad (1)$$

for some $\kappa > 0$. The general solution to (1) can be parameterized as

$$f(t) = \alpha + \beta e^{\kappa t},$$

where $t > 0$, $\alpha > 0$, $\beta < 0$, and $\kappa > 0$. More detailed discussion can be found in Seber and Wild (1989).

According to Diggle *et al.* (2002), there are two possible nonlinear regression models in the longitudinal data setting: the correlated error structures (CES) model and the nonlinear random effects (NLRE) model.

We propose a generalization of mixed effects models for the nonlinear longitudinal data analysis, which provides greater flexibility in specifying correlation structures compared to the CES model. It naturally takes closed form expression in the likelihood function, comparing to the clumsy form of NLRE due to nonlinearity of the random effects.

2 Methodology

The specification of a mixed-effects model is that given random effects b_i the responses y_{ij} are independent and follow a distribution from the exponential family, where the conditional mean of $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})'$ is assumed to satisfy

$$g\{E(\mathbf{y}_i|\mathbf{b}_i)\} = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i. \quad (2)$$

There is a need for developing a flexible model that not only offers natural heterogeneity across subjects but also facilitates joint modeling of mean and covariance matrix, or the freedom of specifying a covariance structure for responses. Denote the random effects component $\mathbf{Z}_i b_i$ in equation (2) by a vector \mathbf{U}_i of length n_i with mean $\mathbf{0}$. This \mathbf{U}_i is not necessarily associated with covariates in a general setting, requiring only the condition that $E(\mathbf{U}_i) = \mathbf{0}$. A similar formulation was presented by Heagerty and Zeger (2001). Diggle *et al.* (1998) proposed a similar structure for a spatial GLMM where the random effects were modeled by a Gaussian random field. Other forms of generalization were also proposed, for example, by Lee and Nelder (2001), where they extended the random effects models to cover a broad class of models. They assumed the random effects

component to have the form $\mathbf{Z}_i \mathbf{L}_i(\rho) \mathbf{r}_i$, where $\mathbf{r}_i \sim \text{MVN}_{n_i}(\mathbf{0}, \Lambda_i)$ and $\mathbf{L}_i(\rho)$ is a $p_i \times n_i$ matrix so that $\mathbf{Z}_i \mathbf{L}_i(\rho) \mathbf{r}_i \sim \text{MVN}_{n_i}(\mathbf{0}, \mathbf{Z}_i \mathbf{L}_i(\rho) \Lambda_i \mathbf{L}_i'(\rho) \mathbf{Z}_i')$. Our setting appears to be a special case of theirs since it leads to \mathbf{r}_i when \mathbf{Z}_i and $\mathbf{L}_i(\rho)$ are both identity matrices. The converse is also true because the random component always results in a vector with mean 0 whatever the form it takes. We further assume $E(\mathbf{U}_i)$ is absorbed into $g\{E(\mathbf{y}_i)\}$ which then is treated to be random. This will lead to a generalization of the model (2). One consequence of such a generalization is that the population heterogeneity can be assumed to come from the different means of the subjects instead of coming from the random effects. That is, a sample of individuals with a common random mean of response is selected from the population and then each of multiple outcomes is measured with certain random noise. This generalization allows flexibility in specifying correlation structure for longitudinal data and in the interpretation of resulting data analysis. Based on the fact that the mean of response is directly assumed random, we will call it the random mean model (RMM).

The random mean model is specified to be very general to furnish flexibility of modelling. For example, the function F may be assumed to be a multivariate t distribution in order to handle outliers, and when a nonlinear function is taken for the mean component, $f(\mathbf{X}_i, \beta)$, it results in a nonlinear RMM.

Following Pourahmadi (1999, 2000) and Pan & Mackenzie (2003), we specify the covariance matrix of F by a modified Cholesky decomposition of Σ^{-1} to achieve a statistically meaningful unconstrained parameterisation of autoregressive coefficients ϕ and the logarithms of the variances, ζ . These parameters may either be modeled, parsimoniously, as functions of covariates.

Pourahmadi (1999, 2000) proposed a reparameterisation of the covariance Σ_i of multivariate normal distribution to remove their constraints and facilitate joint modeling of mean and covariance. The basic idea is based on the Cholesky decomposition of inverse of Σ_i , namely, there exists a unique unit lower triangular matrix \mathbf{T}_i , with 1's as diagonal entries, and a unique diagonal matrix \mathbf{D}_i with positive diagonal entries such that

$$\mathbf{T}_i \Sigma_i \mathbf{T}_i' = \mathbf{D}_i$$

where the below-diagonal entries of \mathbf{T}_i are the negatives of the autoregressive coefficients, ϕ_{ijk} , in $\mu_{ij} = f(\mathbf{X}_{ij}, \beta) + \sum_{k=1}^{j-1} \phi_{ijk}(\mu_{ij} - f(\mathbf{X}_{ij}, \beta))$, the j^{th} random mean component expressed in its predecessors $\mu_{i1}, \dots, \mu_{i(j-1)}$. The diagonal entries of \mathbf{D}_i are the population variances $\sigma_{ij}^2 = \text{var}(\mu_{ij} - f(\mathbf{X}_{ij}, \beta))$.

This set up provides a statistically meaningful way to jointly model mean and covariance. A general model for longitudinal data can be specified as follows,

$$g(\mu_{ij}) = f(\mathbf{x}_{ij}, \beta), \quad \phi_{ijk} = d(\mathbf{z}_{ijk}), \quad \gamma, \quad \log \sigma_{ij}^2 = v(\mathbf{z}_{ij}, \lambda),$$

where $f(\cdot, \cdot)$, $d(\cdot, \cdot)$, $v(\cdot, \cdot)$ are functions, \mathbf{x}_{ij} , \mathbf{z}_{ijk} , \mathbf{z}_{ij} are $p \times 1$, $q_1 \times 1$, $q_2 \times 1$ vectors of covariates, and $\beta = (\beta_1, \dots, \beta_p)'$, $\gamma = (\gamma_1, \dots, \gamma_{q_1})'$, $\lambda = (\lambda_1, \dots, \lambda_{q_2})'$ are parameters corresponding to the mean, dependence and variance, respectively.

In this paper the specifications of variance and autoregressive coefficients take both the approach given in Pan & Mackenzie (2003), i.e. polynomials of time of q_1 , q_2 orders, and some special structures, for example, when $\phi_{ijk} = \gamma$, γ^{j-k} , γ_{j-k} corresponding to compound symmetry, $AR(1)$ and banded correlations. Same as both authors, we choose a information criteria, namely BIC, to select the optimal model, which is defined as,

$$\text{BIC}(p, q_1, q_2) = -\frac{2}{n}\hat{\ell} + (p + q_1 + q_2)\frac{\log n}{n},$$

where $\hat{\ell} = \ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\lambda}})$ is the maximized log-likelihood for the models with the specified degree triple (p, q_1, q_2) . The best triple satisfies

$$(p^*, q_1^*, q_2^*) = \arg \min_{(p, q_1, q_2)} \{\text{BIC}(p, q_1, q_2)\}.$$

REFERENCES

- Diggle, P.J., Heagerty, P., Liang, K.-Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data*, 2nd ed. New York: Oxford University Press.
- Diggle, P.J., Tawn, J.A., Moyeed, R.A. (1998). “Model-based Geostatistics (with discussion)”. *Journal of the Royal Statistical Society, Series B*, 47 (2), 299–350.
- Heagerty, P.J. and Zeger, S.L. (2001). Marginalized multilevel models and likelihood inference. (with discussion). *Statistical Science*, 15, 1–26.
- Lee, Y. and Nelder, J.A. (1996). Hierarchical generalised linear models (with discussion). *Journal of the Royal Statistical Society, Series B*, 58, 619–678.
- Meyers, S.M., Ambler, J.S., Tan, M., Werner, J.C. & Huang, S.S. (1992). Variation of perfluoropropane disappearance after vitrectomy. *Retina* **12**, 359–363.
- Pan, J. and Mackenzie, G. (2003). On modelling mean-covariance structures in longitudinal studies. *Biometrika* **90**, 239–244.
- Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika* **86**, 677–690.
- Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika* **87**, 425–435.
- Seber, G.A.F and Wild C.J. (1989). Nonlinear regression. New York: Wiley.